ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

A Survey of the Security Use Cases in Big Data

Kudakwashe Zvarevashe[1], Mainford Mutandavari[2], Trust Gotora[3]

M Tech Student, Dept of CSE, Jawaharlal Nehru Technological University, Hyderabad, India [1]

M Tech Student, Dept of CSE, Jawaharlal Nehru Technological University, Hyderabad, India [2]

M Tech Student, Dept of Software Engineering, Jawaharlal Nehru Technological University, Hyderabad, India [3]

ABSTRACT: Big data is the collection of large and complex data sets that are difficult to process using on-hand database management tools or traditional data processing applications. The invention of online social networks, smart phones, fine tuning of ubiquitous computing and many other technological advancements have led to the generation of multiple petabytes of both structured, unstructured and semi-structured data. These massive data sets have lead to the birth of some distributed data processing and storage technologies like Apache Hadoop and MongoDB. To tackle the security issues in Hadoop, the Kerberos protocol has been introduced in its second edition. However, this technological movement has created some security loopholes in the processing and storage of the data sets. This paper tries to list some of the directions research on Big Data challenges has taken for the past five years together with their corresponding Use cases.

***KEYWORDS*:** big data, apache Hadoop, MongoDB, Kerberos, NoSQL and social networks.

I. INTRODUCTION

According to SINTF, 90% of the world's data was has been generated over the past two years. The emergence of new and advanced technologies over the past decade has boosted the data consumers' appetite to create, store and consume data [1][2]. CISCO VNI Mobile Focast has highlighted that Asia alone is expected to have a 76% compound annual growth rate for mobile only data [3]. This has since stimulated the desire to address the problems of processing and storage of these vast amounts of data sets. Apache Hadoop and many other technologies have been the knights in shining armour for this problem.

This development has had a positive impact on the way large data sets are being processed together with the storage issues. However, less activities have been done to strengthen the security of the Big data infrastructure together with the data. Some researchers have came up with the Kerberos protocol to handle the security issues in Hadoop but apparently there is a plethora of security issues which range from computations in distributed programming frameworks to data provenance. The phenomenon of large data already exists in the fields of physics, biology, environmental ecology, automatic control and other scientific area. It is also crucial in military, communication finance and many other areas. This clearly qualifies Big data as an information security problem which has a lot of challenges which have to be curbed.

II. RELATED WORK

The growth of big data has raised a number of eyebrows as far as the challenges are concerned. Several authors have discovered a plethora of challenges which include data storage and privacy. Xiaoxue Zhang et al described the storage challenges of Big Data and they analysed them using Social Networks as examples. They further classified the related research issues into the following classifications: small files problem, load balancing, replica consistency and de-duplication. Meiko Johnson also did some work on the privacy issues involved with Big Data. He classified these challenges into the following taxonomy: interaction with individuals, re-identification attacks, probable vs. provable results, targeted identification attacks and economics effects. visualize and understand their algorithm results. Kapil Bakshi et al [9] discussed the architectural considerations for Big data are concluded that despite the different architectures and design decisions, the analytics systems aim for Scale-out, Elasticity and High availability. Sachchidanand Singh et al in[10] described all the concepts of Big data along with the available market solutions used

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

to handle and explore the unstructured large data are discussed. The observations and the results showed that analytics has become an important part for adding value for the social business.

III. CHARACTERISTICS OF BIG DATA

Big data is a term used to describe the collection of large and complex data sets that are difficult to process using on-hand database management tools or traditional data processing applications. Big data spans across seven dimensions which include volume, variety, volume, value ,veracity , volatility and complexity [4].

- Volume: The volume of data here is very huge and is generated from a lot of different devices. The size of the data is usually in terabytes and petabytes. All this data also needs to be encrypted for privacy protection.
- Velocity: This describes the real time attribute found in some of the data sets for example streaming data. The result that misses the appropriate time is usually of little value.
- Variety: Big data consists of a variety of different types of data i.e. structured, unstructured and semi-structured data. The data maybe in the form of blogs, videos, pictures, audio files, location information etc.
- Value: This refers to the complex, advanced, predictive, business analysis and insights associated with the large data sets.
- Veracity: This deals with uncertain or imprecise data. It refers to the noise, biases and abnormality in data. This is where we find out if the data that is being stored and mined is meaningful to the problem being analyzed.
- Volatility: Big Data volatility refers to how long the data is going to be valid and how long it should be stored.
- Complexity: A complex dynamic relationship often exists in Big data. The change of one data might result in the change of more than one set of data triggering a rippling effect

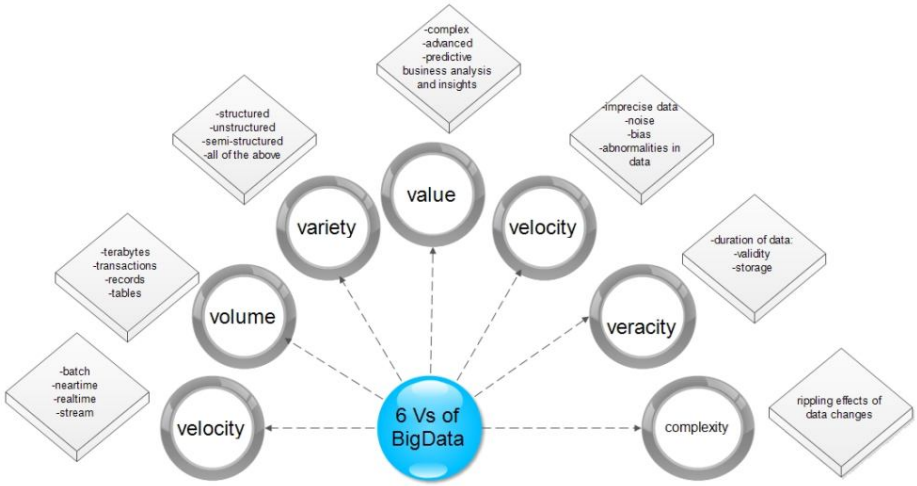

Fig.1:Characteristics of Big Data

IV. THE SECURITY USE CASES

A) Secure Computations in Distributed Programming Frameworks

Distributed programming frameworks use the parallelism concept in computation and storage to process massive amounts of data. The MapReduce framework is a popular example which splits an input file into multiple chunks. In the first phase of MapReduce, a Mapper for each chunk reads the data, performs some computation, and outputs a list of key/value pairs. In the next phase, a Reducer combines the values belonging to each distinct key and outputs the result. There are two major attack prevention measures: securing the mappers and securing the data in the presence of an untrusted mapper [6][5].

Use Cases

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

There is a possibility that untrusted mappers could return wrong results, which will in turn generate incorrect aggregate results. With large data sets, it is almost impossible to identify the fault, resulting in significant damage, especially for scientific and financial computations.
Retailer consumer data is often scrutinised by marketing agencies for targeted advertising or customer-segmenting. These tasks involve highly parallel computations over large data sets, and are particularly suited for MapReduce frameworks such as Hadoop. However, the data mappers may contain intentional or unintentional leakages. For example, a mapper may emit a very unique value by analysing a private record, undermining users' privacy.

B) Security Best Practices for Non-Relational Data Stores
Non-relational data stores have not yet reached security infrastructural maturity. These stores are designed mainly through the use of NoSQL databases. NoSQL Databases were built to tackle different obstacles brought about by the analytics world and hence security was never part of the model at any point of its design stage. Developers using NoSQL databases usually embed security in the middleware. NoSQL databases do not provide any support for enforcing it explicitly in the database. However, clustering aspect of NoSQL databases poses additional challenges to the robustness of such security practices [6][5].

Use Cases
Since Big Data involves large volumes of data which may be structured , semi-structured or unstructured. It is much easier for Companies dealing with big unstructured data sets to benefit in migrating from a traditional relational database to a NoSQL database in terms of accommodating/processing huge volume of data. In general, the security philosophy of NoSQL databases relies in external enforcing mechanisms. To reduce security incidents, the company must review security policies for the middleware adding items to its engine and at the same time toughen NoSQL database itself to match its counterpart RDBs without compromising on its operational features.

C) Secure Data Storage and Transactions Logs
Data and transaction logs are stored in multi-tiered storage media. Manually moving data between tiers gives the IT manager direct control over exactly what data is moved and when. However, as the size of data set has been, and continues to be, growing exponentially, scalability and availability have necessitated auto-tiering for big data storage management. Auto-tiering solutions do not keep track of where the data is stored, which poses new challenges to secure data storage. New mechanisms are imperative to thwart unauthorized access and maintain the 24/7 availability [6][5].

Use Cases
A manufacturer wants to integrate data from different divisions. Some of this data is rarely retrieved, while some divisions constantly utilize the same data pools. An auto-tier storage system will save the manufacturer money by pulling the rarely utilized data to a lower (and cheaper) tier. However, this data may consist in R&D results, not popular but containing critical information. As lower-tier often provides decreased security, the company should study carefully tiering strategies [6].

D) End-Point Input Validation/Filtering
Many big data use cases in enterprise settings require data collection from many sources, such as end-point devices. For example, a security information and event management system (SIEM) may collect event logs from millions of hardware devices and software applications in an enterprise network. A key challenge in the data collection process is input validation: how can we trust the data? How can we validate that a source of input data is not malicious and how can we filter malicious input from our collection? Input validation and filtering is a daunting challenge posed by untrusted input sources, especially with the bring your own device (BYOD) model [6][5].

Use Cases
Both data retrieved from weather sensors and feedback votes sent by an iPhone application share a similar validation problem. A motivated adversary may be able to create "rogue" virtual sensors, or spoof iPhone IDs to rig the results. This is further complicated by the amount of data collected, which may exceed millions of readings/votes. To perform these tasks effectively, algorithms need to be created to validate the input for large data sets.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

E) Real-time Security/Compliance Monitoring

Real-time security monitoring has always been a challenge, given the number of alerts generated by (security) devices. These alerts (correlated or not) lead to many false positives, which are mostly ignored or simply "clicked away," as humans cannot cope with the shear amount. This problem might even increase with big data, given the volume and velocity of data streams. However, big data technologies might also provide an opportunity, in the sense that these technologies do allow for fast processing and analytics of different types of data. Which in its turn can be used to provide, for instance, real-time anomaly detection based on scalable security analytics [6][5].

Use Cases

Most industries and government (agencies) will benefit from real-time security analytics, although the use cases may differ. There are use cases which are common, like, "Who is accessing which data from which resource at what time"; "Are we under attack?" or "Do we have a breach of compliance standard C because of action A?" These are not really new, but the difference is that we have more data at our disposal to make faster and better decisions (e.g., less false positives) in that regard. However, new use cases can be defined or we can redefine existing use cases in lieu of big data. For example, the health industry largely benefits from big data technologies, potentially saving billions to the tax-payer, becoming more accurate with the payment of claims and reducing the fraud related to claims. However, at the same time, the records stored may be extremely sensitive and have to be compliant with HIPAA or regional/local regulations, which call for careful protection of that same data. Detecting in real-time the anomalous retrieval of personal information, intentional or unintentional, allows the health care provider to timely repair the damage created and to prevent further misuse.

F) Scalable and Composable Privacy-Preserving Data Mining and Analytics

Big data can be seen as a troubling manifestation of Big Brother by potentially enabling invasions of privacy, invasive marketing, decreased civil freedoms, and increase state and corporate control. A recent analysis of how companies are leveraging data analytics for marketing purposes identified an example of how a retailer was able to identify that a teenager was pregnant before her father knew. Similarly, anonymizing data for analytics is not enough to maintain user privacy. For example, AOL released anonymized search logs for academic purposes, but users were easily identified by their searchers. Netflix faced a similar problem when users of their anonymized data set were identified by correlating their Netflix movie scores with IMDB scores. Therefore, it is important to establish guidelines and recommendations for preventing inadvertent privacy disclosures [6][5].

Use Cases

User data collected by companies and government agencies are constantly mined and analysed by inside analysts and also potentially outside contractors or business partners. A malicious insider or untrusted partner can abuse these datasets and extract private information from customers.

Similarly, intelligence agencies require the collection of vast amounts of data. The data sources are numerous and may include chat-rooms, personal blogs and network routers. Most collected data is, however, innocent in nature, need not be retained, and anonymity preserved.
Robust and scalable privacy preserving mining algorithms will increase the chances of collecting relevant information to increase user safety.

G) Cryptographically Enforced Access Control and Secure Communication

To ensure that the most sensitive private data is end-to-end secure and only accessible to the authorized entities, data has to be encrypted based on access control policies. Specific research in this area such as attribute-based encryption (ABE) has to be made richer, more efficient, and scalable. To ensure authentication, agreement and fairness among the distributed entities, a cryptographically secure communication framework has to be implemented [6][5].

Use Cases

Sensitive data is routinely stored unencrypted in the cloud. The main problem to encrypt data, especially large data sets, is the all-or-nothing retrieval policy of encrypted data, disallowing users to easily perform fine grained actions such as sharing records or searches. ABE alleviates this problem by utilizing a public key cryptosystem where attributes related

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

to the data encrypted serve to unlock the keys. On the other hand, we have unencrypted less sensitive data as well, such as data useful for analytics. Such data has to be communicated in a secure and agreed-upon way using a cryptographically secure communication framework.

H) Granular Access Control

The security property that matters from the perspective of access control is secrecy—preventing access to data by people that should not have access. The problem with course-grained access mechanisms is that data that could otherwise be shared is often swept into a more restrictive category to guarantee sound security. Granular access control gives data managers a scalpel instead of a sword to share data as much as possible without compromising secrecy [6][5].

Use Cases

Big data analysis and cloud computing are increasingly focused on handling diverse data sets, both in terms of variety of schemas and variety of security requirements. Legal and policy restrictions on data come from numerous sources. The Sarbanes-Oxley Act levees requirements to protect corporate financial information, and the Health Insurance Portability and Accountability Act includes numerous restrictions on sharing personal health records. Executive Order 13526 outlines an elaborate system of protecting national security information.

Privacy policies, sharing agreements, and corporate policy also impose requirements on data handling. Managing this plethora of restrictions has so far resulted in increased costs for developing applications and a walled garden approach in which few people can participate in the analysis. Granular access control is necessary for analytical systems to adapt to this increasingly complex security environment.

I) Granular Audits

With real-time security monitoring , we try to be notified at the moment an attack takes place. In reality, this will not always be the case (e.g., new attacks, missed true positives). In order to get to the bottom of a missed attack, we need audit information. This is not only relevant because we want to understand what happened and what went wrong, but also because compliance, regulation and forensics reasons. In that regard, auditing is not something new, but the scope and granularity might be different. For example, we have to deal with more data objects, which probably are (but not necessarily) distributed [6][5].

Use Cases

Compliance requirements (e.g., HIPAA, PCI, Sarbanes-Oxley) require financial firms to provide granular auditing records. Additionally, the loss of records containing private information is estimated at $200/record. Legal action – depending on the geographic region – might follow in case of a data breach. Key personnel at financial institutions require access to large data sets containing PI, such as SSN. Marketing firms want access, for instance, to personal social media information to optimize their customer-centric approach regarding online ads.

J) Data Provenance

Provenance metadata will grow in complexity due to large provenance graphs generated from provenance-enabled programming environments in big data applications. Analysis of such large provenance graphs to detect metadata dependencies for security/confidentiality applications is computationally intensive [6][5].

Use Cases

Several key security applications require the history of a digital record – such as details about its creation. Examples include detecting insider trading for financial companies or to determine the accuracy of the data source for research investigations. These security assessments are time sensitive in nature, and require fast algorithms to handle the provenance metadata containing this information. In addition, data provenance complements audit logs for compliance requirements, such as PCI or Sarbanes-Oxley.

These security issues can be classified into four distinct aspects of the Big Data Ecosystem as shown in Fig. 2 below.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

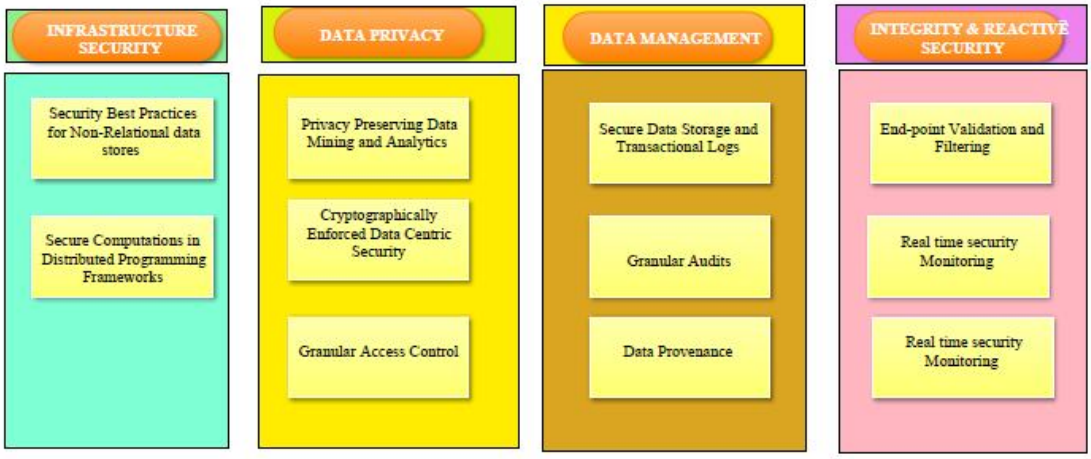

Fig. 2:Taxonomy of Big Data Challenges

V. POSSIBLE SOLUTIONS TO BIG DATA CHALLENGES

i)Kerberos

Kerberos is a system for authenticating users and services on a network. The goal of Kerberos is to validate nodes to ensure unwanted copies of data or unwanted queries are run to copy data. In traditional IT, this may seem far-fetched, but in cloud and virtual environments with thousands of nodes in a cluster, this is both easy and difficult to detect. Kerberos provides a means of authenticating nodes before they are allowed to participate on the cluster. This solution has already been implemented in the latest edition of Apache Hadoop.

ii)TLS

Transport Layer Security (TLS) is a protocol that ensures privacy between communicating protocol and their users on the Internet. The goal here is to keep communications private. Built-in Hadoop capabilities provide for secure client-to-Web application-layer communication, but not internode communication, nor Mapper-Reducer output before it is passed to the application. TLS provides a mechanism for secure communication between all nodes and services, and scales as nodes are added to the cluster as illustrated in Fig. 3.

iii)File Layer Encryption

The idea is to protect the contents of the cluster from those who administer it. In the event an administrator wants to inspect data files, either through a basic text editor or by taking a snapshot and examining the archive, encryption keeps data safe. File-layer encryption is transparent to the database, and scales up as new nodes are added to the cluster. Granted, the encryption keys must be kept secure for this feature to be effective.

iv)Key Management

As an extension of file-layer encryption, key management is critical to successful encryption deployment. All too often when we review cloud and virtual security systems, we find that administrators keep encryption keys stored unprotected on the disk. Encryption keys must be readily available when a cluster restarts; otherwise, data is inaccessible, and leaving keys in the open was the only way they knew how to ensure safe system restarts. Most central key management systems provide for key negotiation on restart, and still keep keys safe.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

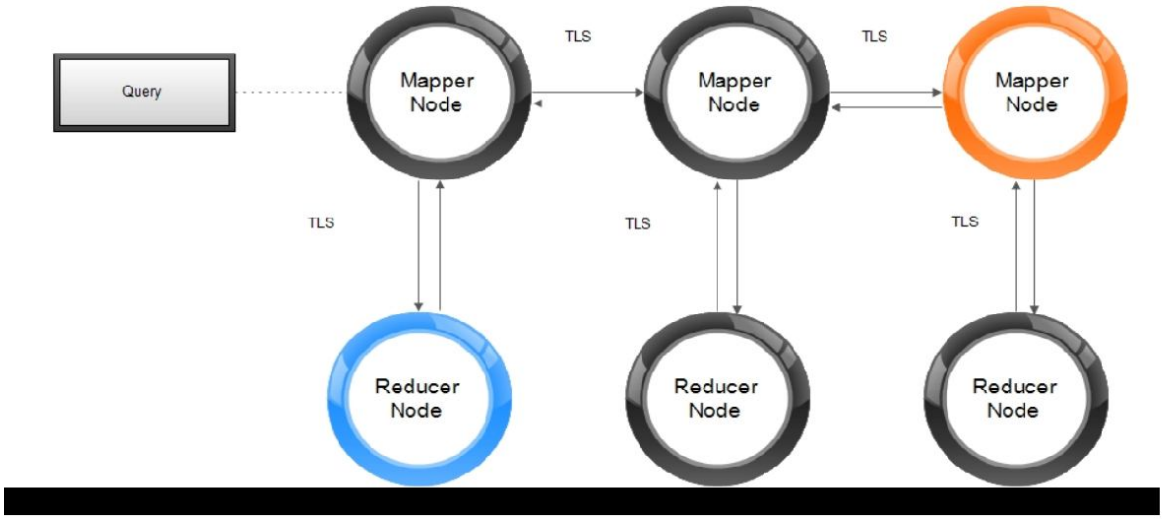

Fig. 3:Illustration of TLS internode proposed implementation

v)Deployment Validation

The goal here it to ensure that all nodes in the cluster are properly patched, properly configured, and run the correct (i.e., not hacked) copies of software. It's easy to miss things when you have thousands of nodes running, so automated node validation makes basic node security much easier. Scripts to install patches and set and test configuration settings are an easy way to verify no nodes enter the cluster without proper setup and security. This can also be linked into Kerberos authentication and external application validation (security as a service) offerings.

VI. CONCLUSION AND FUTURE WORK

This paper has exposed the major security problems that need to be addressed in Big Data processing and storage. Some researchers have brought about the use of encryption together with Kerberos protocol in order to make the data more secure. However, these security and privacy issues come in different forms such that Kerberos might not be enough to fully secure the data. During Map-Reduce framework in Hadoop, mapper nodes process a given set of data and saves the intermediary data within their local files. The reducer nodes will then copy this data from the mapper nodes and later on aggregate it to produce the overall result. We would like to introduce an additional central node which connects with both the mapper and the reducer nodes. The intermediary data will then be stored in this node instead of the mapper nodes' local file system. A perimeter defence mechanism will then be used to monitor all the traffic going into and out of the node to secure the data.

REFERENCES

1. http://www.sintef.no
2. http://www.sciencedaily.com/releases/2013/05/130522085217.html
3. http://www.cisco.com/web/solutions/sp/vni/vni_mobile_forecast_highlight/index.html
4. Xiaoxue Zhang, Feng Xu, "Survey of Research on Big Data Storage", 2013 12th International Symposium on Distributed Computing and Applications to Business, Engineering & Science
5. Top Ten Big Data Security And Privacy Challenges CLOUD SECURITY ALLIANCE https://cloudsecurityalliance.org/
6. https://cloudsecurityalliance.org/media/news/csa-releases-the-expanded-top-ten-big-data-security-privacy-challenges/
7. http://www.darkreading.com/views/dont-get-ha-duped-by-big-data-security-p/240144305
8. Sachchidanand Singh, Nirmala Singh, "Big Data Analytics", 2012 International Conference on Communication, Information & Computing Technology (ICCICT), Oct. 19-20, Mumbai, India
9. Kapil Bakshi, "Considerations for Big Data: Architecture and Approach", IEEE Aerospace conference 2012

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

10. Sachchidanand Singh, Nirmala Singh, "Big Data Analytics", IEEE, international Conference on Communication, Information & Computing *Technology (ICCICT)*, Oct. 19-20, 2012.

BIOGRAPHY

Kudakwashe Zvarevashe: Attained his BSc degree in Information Systems at MSU, Zimbabwe in 2010. He is currently doing M Tech IT final year at JNTUH, India. He is a HIT staff development research fellow. His research interests are in the area of big data, information security, cloud computing and web services.

Mainford Mutandavari: Attained his BSC Degree in Computer Science from MSU, Zimbabwe in 2010. Currently he is studying towards M.Tech CSE at JNTUH, India. . He is a HIT staff development research fellow. His research interests are in Cloud Computing ,Big Data ,Web Services , Networking and Information Systems.

Tatenda Trust Gotora: Attained his BSc degree in Computer Science at MSU, Zimbabwe in 2011. He is currently doing M Tech SE final year at JNTUH, India. He is a HIT staff development research fellow. His research interests are in the area of mobile computing, big data, information security and web services.