ISSN No: - 2456 - 2165

Big Data - A Nightmare to Data Scientist

B. Lakshmi
Asst. Professor, Department of Computer Applications,
V.R.S.E.C, Vijayawada -7, Andhra Pradesh, India

K. Anji Reddy
Head of the Department, Dept. of Computer Applications,
V.R.S.E.C, Vijayawada -7, Andhra Pradesh, India

Abstract **- The rise of social media and mobile devices involves rapid growth in data generation as a result data management has become increasingly challenging. The term Big Data is coined for large scale of data, which contains information including audio and video files. As data prone to rapid changes, traditional database management techniques are not capable to maintain such large varying data. More over data analysis has become a nightmare to data scientists. New analytics tools are emerging into IT world with a perspective to analyze the big data. Hadoop (IOP) and Big Insights together provide a software platform for visualizing, discovering, and analyzing data from disparate sources. This paper spots a light on big data characteristics and comparison between old and new architectures of data management. This survey paper concludes with a discussion of Hadoop as a solution to big data and promising future directions.**

Index Terms **–** *Big Data, Zetta Byte, SaaS, Variability.*

I. INTRODUCTION

The problem of working with big data which exceeds the computing power is not new; this type of computing has greatly wide spreading in recent years. Big data is a term used for large scale of data that comes in all shapes and sizes which prone to rapid change and make them difficult to be captured.

An exact definition of big data is difficult to nail down because projects, vendors, practitioners, and business professionals use it quite differently. With that in mind, big data is:

- large datasets
- the set of computing strategies and technologies that are used to maintain large datasets

Large dataset means a dataset too large to reasonably process or store with traditional tooling or on a single computer. This means that the common scale of big datasets is constantly shifting and may vary significantly from organization to organization.

Figure: 1. Big Data

II. CATEGORIES OF BIG DATA

Big data is broadly classified into three categories:

a. Structured
b. Unstructured
c. Semi-structured

A. Structured

Data that can be stored in a fixed format is called structured data. There are many database management techniques to maintain and analyze this type of data. When size of this data increased to large extent, challenges involved in its discovery, management and processing are also increased. Tabular data stored in RDBMs is example of a structured data.

Example:

RELATION NAME	COLUMN NAME	DATA TYPE	SIZE	CONSTRAINT
EVENTMANAGER:				
EVENTMANAGER	UNAME	VARCHAR2	30	PRIMARY KEY
EVENTMANAGER	PSW	VARCHAR2	30	NOT NULL
EVENTMANAGER	NAME	VARCHAR2	50	
EVENTMANAGER	CNAME	VARCHAR2	50	
EVENTMANAGER	ADDRESS	VARCHAR2	50	
EVENTMANAGER	EID	VARCHAR2	40	
EVENTMANAGER	CNO	NUMBER	10	
EVENTMANAGER	STATUS	NUMBER	10	CHECK IN ('PENDING', 'ACTIVE', 'DEACTIVE')

Figure: 2. Example of Structured Data

ISSN No: - 2456 - 2165

A. Unstructured

Data with unknown format or structure is called unstructured data. Processing of data and deriving values become major challenges of large datasets of un-structured. Heterogeneous data sets containing a set of simple text files, images, videos etc are best examples of un-structured data.

Example: Result of Search Engine

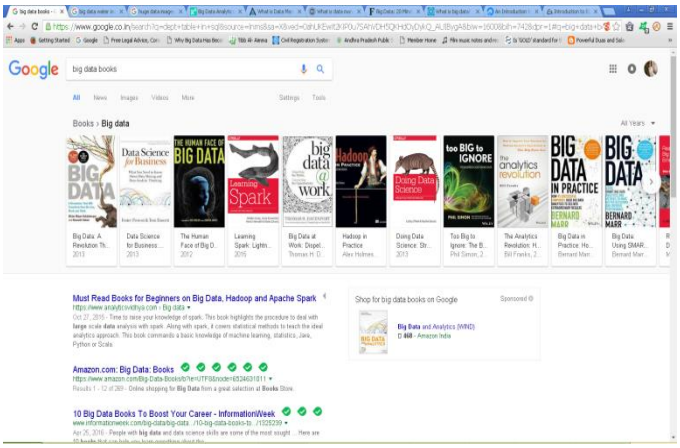

Figure: 3. Examples of Un-Structured Data

B. Semi-structured

A combination of both structured and un-structured data is called Semi-structured data. The resulted output of semi-structured data has particular format but it is really not defined with. E.g. a table definition of RDBMS table and data stored in XML file.

Example:

```
<breakfast_menu>
  <food>
    <name>Belgian Waffles</name>
    <price>$5.95</price>
    <description>
      Two of our famous Belgian Waffles with plenty of real maple syrup
    </description>
    <calories>650</calories>
  </food>
  <food>
    <name>Strawberry Belgian Waffles</name>
    <price>$7.95</price>
    <description>
      Light Belgian waffles covered with strawberries and whipped cream
    </description>
    <calories>900</calories>
  </food>
  <food>
    <name>Berry-Berry Belgian Waffles</name>
    <price>$8.95</price>
    <description>
      Light Belgian waffles covered with an assortment of fresh berries and whipped cream
    </description>
    <calories>900</calories>
  </food>
  <food>
    <name>French Toast</name>
    <price>$4.50</price>
    <description>
      Thick slices made from our homemade sourdough bread
    </description>
    <calories>600</calories>
  </food>
  <food>
    <name>Homestyle Breakfast</name>
    <price>$6.95</price>
    <description>
      Two eggs, bacon or sausage, toast, and our ever-popular hash browns
    </description>
    <calories>950</calories>
  </food>
</breakfast_menu>
```

Figure: 4. Example of Semi-Structured Data

III. CHARACTERISTICS OF BIG DATA

The basic requirements for working with Big Data are the same as the requirements for working with datasets of any size. However, the massive scale, the speed of ingesting and processing, and the characteristics of the data that must be dealt with at each stage of the process present significant new challenges when designing solutions.

The goal of most Big Data systems is to surface insights and connections from large volumes of heterogeneous data that would not be possible using conventional methods.

A. Volume

The sheer scale of the information processed helps define Big Data systems. Data at rest in big data can be orders of larger magnitudes than traditional datasets, which requires more professional techniques to store. Hadoop distributed file system and new Cluster management algorithms become more prominent as they break tasks into smaller pieces and become important.

B. Velocity

Data in motion (Stream data with, millisecond to respond) - the speed with which information moves through the system. Data is frequently flowing into the system from multiple sources example; audio, video data is constantly being added, massaged, processed, and analyzed in social media. These systems require robust and reliable systems to guard against failures and to maintain data integrity and security against networks.

C. Variety

Data in many forms (Like structured, unstructured, text, multimedia) - The formats and types of media can vary significantly as well. Big data is set of different formats of data like images, video files, audio, text files and structured logs, etc.

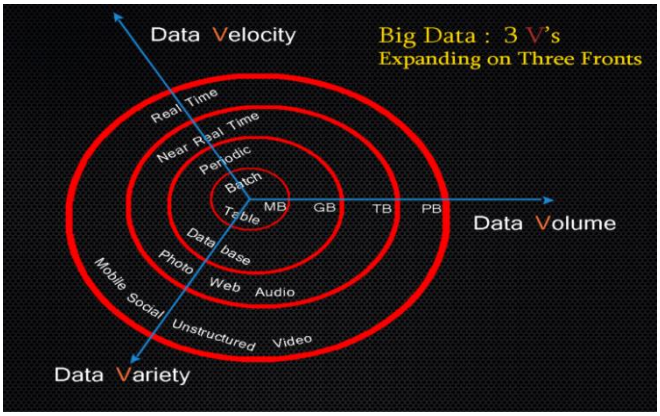

Figure: 5. Big Data as 3Vs

ISSN No: - 2456 - 2165

D. Other Characteristics

Various individuals and organizations have suggested expanding the original three Vs, though these proposals have tended to describe challenges rather than qualities of Big Data. Some common additions are:

- ***Veracity***
  The variety of sources and the complexity of the processing can lead to challenges in evaluating the quality of the data (and consequently, the quality of the resulting analysis).

- ***Variability***
  Variation in the data leads to wide variation in quality. Additional resources may be needed to identify, process, or filter low quality data to make it more useful.

- ***Value***
  The ultimate challenge of Big Data is delivering value. Sometimes, the systems and processes in place are complex enough that using the data and extracting actual value can become difficult.

IV. ARCHITECTURE

Traditional Database management systems used client server architecture to process data.

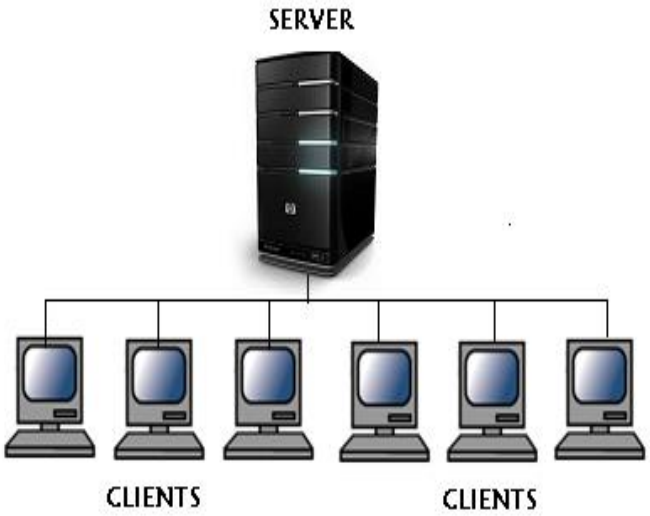

Figure: 6. Client Server Architecture

Normal data like small in size will be processed easily for decision making but big data with larger datasets requires parallel processing. So Big data is processed with Master Slave architecture

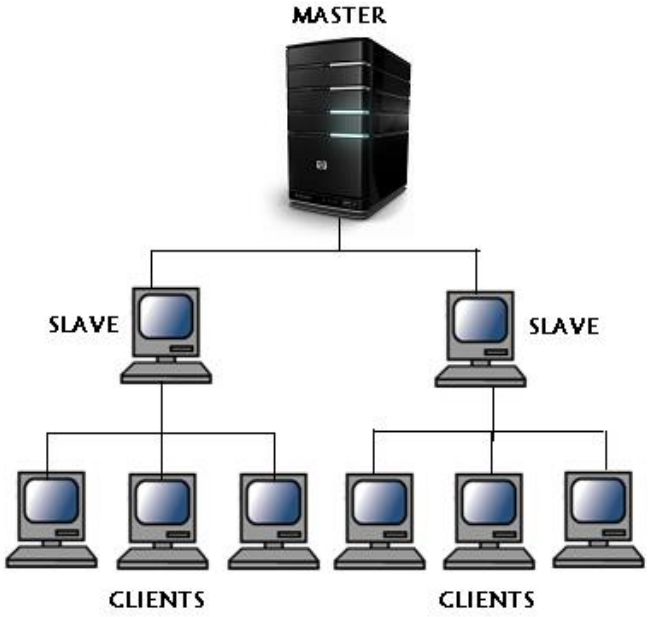

Figure: 7. Master/Slave Architecture

A. Hadoop As A Solution

Apache Hadoop is an open-source Programming framework used to support distributed storage and processing of large data sets using the MAPREDUCE programming model a software framework where an application break down into various parts.

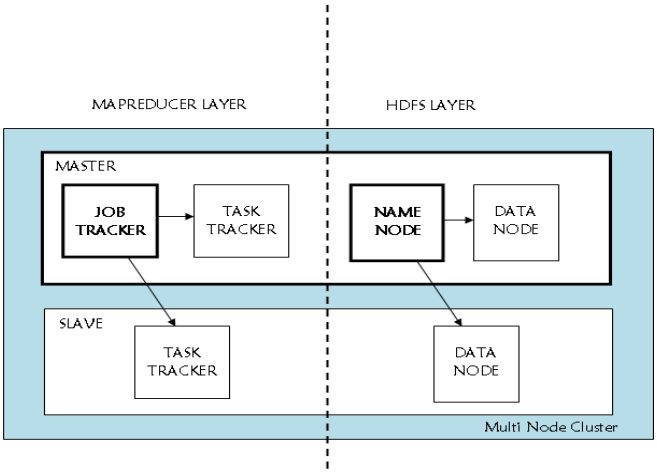

Figure: 8. Hadoop Architecture

B. Main Components of HADOOP

1. HDFS (digital data)
2. MR(Map Reduce-write business logic to process (written in core java))

3. SQOOP(SQL+HADOOP: Can export or import SQL data in Hadoop or vice versa)
4. HIVE(Data warehouse)
5. HBASE(NO SQL components)
6. OOZIE (workflow)
7. FLUME(contnous streaming data like twitter, facebook etc)
8. PIG (predefinedcomponents used for processing like MapReduce)

V. HDFS ARCHITECTURE

The term HDFS called Hadoop distributed file system and it is the fault tolerant storage component of Hadoop framework.

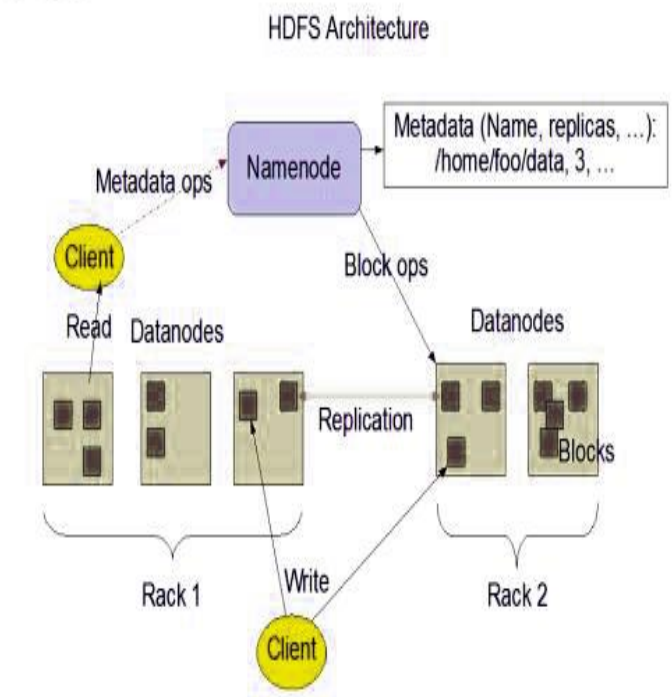

Figure: 9. HDFS Architecture

HDFS can store large scale of information, upgrade incrementally and checks the system periodically to project data from loss and maintains integrity. Hadoop uses Master slave architecture which creates clusters nothing but a set of computer and coordinates work among them. This master slave system prevents data loss and interrupting work as it maintains the work load with other machines in cluster. HDFS replicate the pieces of incoming files, called "blocks" and stores across multiple machines in the cluster and to different servers.

A. Map Reducer Architecture

Processing part of the Bigdata is done by MapReduce (Business logic).

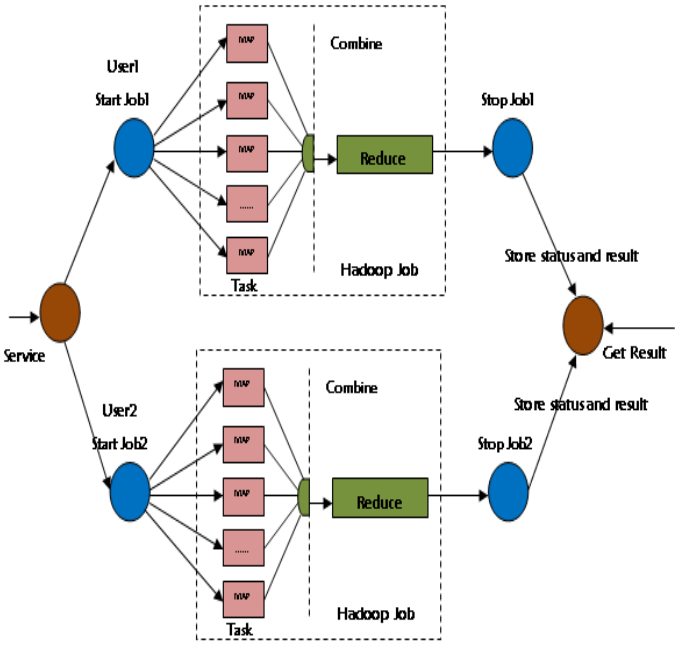

Figure: 10. MapReducer Architecture

MapReduce framework is the processing backbone of hadoop architecture. The framework divides the specifications (processing logic or business logic) of operations which are used to process large datasets, and run them in parallel.

In Hadoop, specifications are written as MapReduce jobs in Java. Operations can also be written in Hive and pig. HDFS stores the all the combined results of MapReducer. The two major functions of MapReducer are MAP and REDUCE.

MAP – This function receives input from user, processes the input using specifications of operations and generates an intermediate set of output pairs.

REDUCE – This function merges all the intermediate values generated by the MAP.

Finally the stored status and results are stored in HDFS.

B. Key Points of Hadoop

Hadoop is one of the emerging technologies in IT industry to overcome the challenges in processing of Big, rapidly changing and venerable datasets of Big Data. It is open source framework and has many advantages. The five Key points of Hadoop are:

1. Built on Java Technology
2. Cost effective
3. Fault tolerant
4. Scalability and capacity increased by adding nodes
5. Industry chosen technology for performing analytics on un-structured data

C. Cloudera Distribution for Hadoop Vs Ibm Infosphere Biginsights

Cloudera Distribution for Hadoop is the world's popular, most complete, tested, and popular distribution of Apache Hadoop and related domains. CDH is 100% Apache-licensed open source and is the only Hadoop solution to offer unified batch processing, interactive SQL, and interactive search, and role-based access controls. More enterprises have downloaded CDH than all other such distributions combined.

IBM BigInsights delivers a rich set of advanced analytics capabilities that allows enterprises to analyze massive volumes of structured and unstructured data in its native format. The software combines open source Apache Hadoop with IBM innovations including sophisticated text analytics, IBM BigSheets for data exploration, IBM Big SQL for SQL access to data in Hadoop, and a range of performance, security and administrative features. The result is a cost-effective and user-friendly solution for complex, big data analytics.

D. Infosphere Biginsights

InfoSphere BigInsights v3.0 is a software platform designed to help organizations discover and analyze business insights hidden in large volumes of a diverse range of data.

Examples of such data include log records, online shopping, click streams, social media data, news feeds, and electronic sensor output.

To help firms derive value from such data in an efficient manner, BigInsights incorporates several open source projects (including Apache™ Hadoop™) and a number of IBM-developed technologies. Basic word count problem is solved through BigInsights in the following manner:

Figure: 11. BigInsights v3.0

Click Start BigInsights to start all required services.

Figure: 12. Icon of BigInsights

To verify that all required BigInsights services are up and running from a terminal window, issue this command: $BIGINSIGHTS_HOME/bin/status.sh. hdm, zookeeper, hadoop, catalog, hive, bigsql, oozie, comsole, and httpfs all the components started successfully.

```
[INFO] Progress - Status hbase
[INFO] @bivm.ibm.com - hbase-master(active) started, pid 8407
[INFO] @bivm.ibm.com - hbase-regionserver started, pid 8573
[INFO] Deployer - hbase service started
[INFO] Progress - 45%
[INFO] Progress - Status hive
[INFO] @bivm.ibm.com - hive-web-interface started, pid 9746
[INFO] @bivm.ibm.com - hive-server2 started, pid 9997
[INFO] Progress - 55%
[INFO] Progress - Status bigsql
[INFO] @bivm.ibm.com - bigsql-server started, pid 10622
[INFO] @bivm.ibm.com - head-node started, pid 15033
[INFO] @bivm.ibm.com - work-node started, pid 15028
[INFO] @bivm.ibm.com - scheduler started, pid 16826
[INFO] Progress - 64%
```

Figure: 13. Progress – Status of components

To find the word count: hadoop fs –ls WordCount_output

ISSN No: - 2456 - 2165

```
biadmin@bivm:~> hadoop fs -ls WordCount_output
Found 3 items
-rw-r--r--   1 biadmin biadmin          0 2014-07-02 18:42 WordCount_output/_SUCCESS
drwx--x--x   - biadmin biadmin          0 2014-07-02 18:42 WordCount_output/_logs
-rw-r--r--   1 biadmin biadmin      24069 2014-07-02 18:42 WordCount_output/part-r-00000
```

Figure: 14. Running the command

To view the contents of part-r-0000 file: hadoop fs –cat WordCount_output/*00

Figure: 15. Partial output

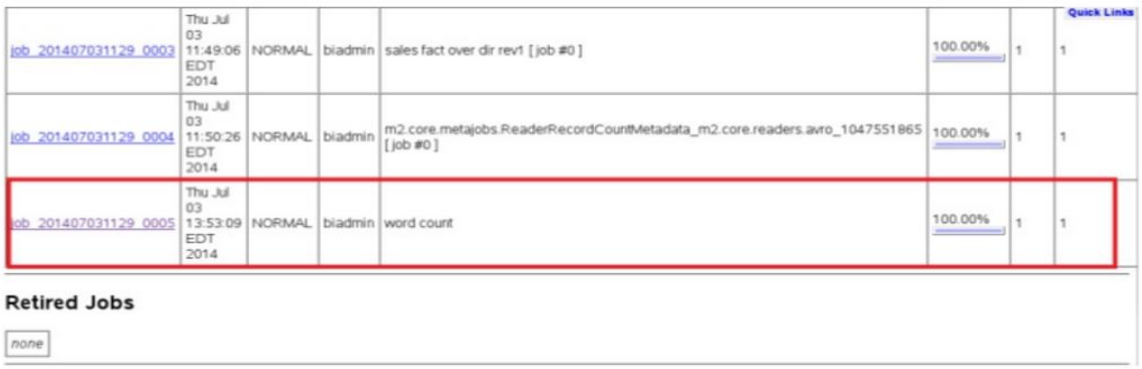

Figure: 16. Execution of word count

VI. CONCLUSION

In this paper the categories of Big Data were presented. Also, the characteristics of Big Data to deal with were discussed. Hadoop was explained as a solution because it enables applications to work with thousands of nodes and petabytes of data in a highly parallel, cost effective manner. This paper also covered the basic word count problem through InfoSphere

BigInsights which is an analytics platform for Big Data. The other side of the coin states that Today's threat environment imposes the three Vs of big data: volume, variety, and velocity. Each of these is increasing at an astonishing rate and has required a swing in how security vendors are going to manage these threats.

REFERENCES

[1] "Implementing Virtual Private Database Using Security Policies", H. Lakshmi, IJEERT, Volume 2, Issue 5, August 2014, PP 146-152.

[2] "Preserving Privacy using Column Masking and Data Encryption Techniques", B. Lakshmi, H. Ravindra Babu, IJCSE, Volume-4, Issue-6,E-ISSN: 2347-2693.

[3] Comparisons of Relational Databases with Big Data: a Teaching Approach, Ali Salehnia, South Dakota State University Brookings, SD 57007

[4] "Big Data: Volume, Velocity, Variability, Variety", http://nosql.mypopescu.com/post/6361838342/bigdata-volume-velocity-variability-variety, Accessed April 2015.

[5] B. Wiederhold, "18 essential Hadoop tools for crunching big data", Network World, www.googletagmanager.com/ns.html, Accessed April, 2015.

[6] A. K. Zaki, "NoSQL Database: New Millennium Database for Big Data, Big Users, Coud Computing and Its Security Challenges," http://esatjournals.org/Volumes/IJRET/2014V03/I15/IJRET20140315080.pdf, Accessed May 2015.

[7] L. Arthur, "What is Big Data", http://www.forbes.com/sites/liasaarthur/2013/08/15/what-is-big-data/, Accessed May 2015.

[8] S. Penchikala, "Virtual Panel: Security Considerations in Accessing NoSQL Databases‖, Nov. 2011. http://www.infoq.com/articles/nosql-data-security-virtual-panel., Accessed May 2015.

[9] Oracle Databases from web: http://www.oracle.com/us/products/database/overview/ ndex.h tml, Accessed May 2015.

[10] Relational Database Management System (RDBMS) vs noSQL.http://openproceedings.org/html/pages/2015_edbt.html, Accessed April 2015.

[11] "Relational -database-management-system-rdbms-vs-nosql/,http://www.loginradius.com/engineering/relational-database-management-system-rdbms-vs-nosql/ Accessed April 2015.

[12] B. Lakshmi, K. Parish Venkata Kumar, A. Shahnaz Banu and K. Anji Reddy, "Data Confidentiality and Loss Prevention using Virtual Private Database."

[13] M. Ramachandran "Relational Vs Non-Relational databases", http://bigdata-madesimple.com/relational-vs-non-relational-databases, Accessed May 2015.

[14] L. P. Issac "SQL vs NoSQL Database Differences Explained with few Example DB", http://www.thegeekstuff.com/2014/01/sql-vs-nosql-db/, Accessed May 2015.

[15] Sherpa Software. "Structured and Unstructured Data: What is It? http://www.sherpasoftware.com/blog/structured-and-unstructured-data-what-is-it/, Accessed May 2015.

[16] F. Chang, et al. "Bigtable: A distributed storage system for structured data.", ACM Transactions on Computer Systems (TOCS) 26.2 (2008): 4.

[17] D. Gosain, "A survey and comparison of relational and non-Relational Databases‖. IJERT,Vol 1,Issue 6,2012.

[18] L. Okman, , N. Gal-Oz, Y. Gonen, E. Gudes, and J. Abramov, , "Security Issues in NoSQL Databases," Trust, Security and Privacy in Computing and Communications (TrustCom), 2

[19] IEEE 10th International Conference on , vol., no., pp.541-547, 16-18 Nov. 2011 doi: 10.1109/TrustCom.2011.70

[20] Find cloud security alliance https://cloudsecurityalliance.org/research/big-data/8

[21] B. Sullivan, "NoSQL, But Even Less Security‖, 2011. http://blogs.adobe.com/asset/files/2011/04/NoSQL-But-Even- Less-Security.pdf. Accessed April 2015.

[22] Find Source: www.couchbase.com/why-nosql/nosql-database.

[23] K. Madia, "The Five Most Interesting Talks at Black Hat, DEF CON and BSides", https://securityintelligence.com/five-steps-for-better-security-analytics-in-2015, Accessed August 2015.

[24] "The Four Pillars of Big Data for the CMO", http://www.certona.com/the-four-pillars-of-big-data-for-the-cmo/, Accessed August 2015.

[25] Unlocking the Magic of Unstructured Content", *IDS White Paper*, http://inmoncif.com/registration/whitepapers/unlocking-final.pdf, Accessed August 2015.

[26] Z.Liu, B. .Jiangz and J. Heer, "imMens: Real-time Visual Querying of Big Data" Eurographics Conference on Visualization (EuroVis) 2013 Volume 32 (2013), Number 3.

[27] S. Durbin**, "**A CEOs Guide to Big Data Security", http://www.infosecurity-magazine.com/view/32736/a-ceos-guide-to-big-data-security/ Accessed May 2015.E. Weindruch"Big Data brings privacy, security concerns",http://www.bizjournals.com/charlotte/print-

[28] edition/2013/09/20/big-data-brings-privacy-security.html?page=all, Accessed May 2015.

[29] "Enhancing Big Data Security", www.advantech.com, Accessed May 2015. H. Chen, R. H. L. Chiang, V. C. Storey, "Business Intelligence and Analytics: From Bog Data to Big impact", *MIS Quarterly*, Special Issue: Business Intelligence Research.

[30] **"**Holistic Approach Needed for Big Data Security"**,** http://www.theiia.org/intAuditor/feature-articles/2013/february/holistic-approach-needed-for-big-data-security/, Accessed May 2015.

[31] J. Hurwitz, A. Nugent, F. Halper, and M. Kaufman, "Security Considerations with Big Data",

http://www.dummies.com/how-to/content/security-considerations-with-big-data.html, Accessed May 2015.
[32] C. White, "Using Big Data for Smarter Decision Making", BI Research, July 2011.
[33] "Addressing Big Data Security Challenges: The Right Tools for Smart Protection", Trend Micro.
[34] S. Grimes, "Unstructured Data and the 80 Percent Rule", http://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/, Accessed May 2015.