International Journal of Multidisciplinary Research and Development

Volume :2, Issue :4, 580-585
April 2015
www.allsubjectjournal.com
e-ISSN: 2349-4182
p-ISSN: 2349-5979
Impact Factor: 3.762

R.Saranya
M.Phil Full Time Research Scholar, Department of computer science, Vivekananda College of Arts and Sciences For Women, Namakkal, TamilNadu, India

V.P.MuthuKumar
Department of computer science and application, Assistant Professor, Vivekananda College of Arts and Sciences For Women, Namakkal, TamilNadu, India.

Correspondence:
R.Saranya
M.Phil Full Time Research Scholar, Department of computer science, Vivekananda College of Arts and Sciences For Women, Namakkal, TamilNadu, India

Security issues associated with big data in cloud computing

R.Saranya, V.P.MuthuKumar

Abstract
The security issues for cloud computing, Big data, Map Reduce and Hadoop environment. It main focus is on security issues in cloud computing that are associated with big data. Big data applications are a great benefit to organizations, business, companies and many large scale and small scale industries. We also discuss various possible solutions for the issue in cloud computing security and Hadoop. Cloud computing security is developing at a rapid pace which includes computer security, network security, information security, and data privacy. Cloud computing plays a very vital role in protecting data, applications and the related infrastructure with the help of policies, technologies, controls, and big data tools**.** Moreover, cloud computing, big data and its applications, advantages are likely to represent the most promising new frontiers in science.

Keywords: Cloud Computing, Big Data, Hadoop,Map Reduce, HDFS (Hadoop Distributed File System)

1. Introduction

Big data is the term for data sets so large and complicated that it becomes difficult to process using traditional data management tools or processing applications.The data and to identify patterns it is very important to securely store,manage and share large amounts of complex data.Cloud comes with an explicit security challenge, i.e. the data owner might not have any control of where the data is placed.Apache's Hadoop distributed file system (HDFS) is evolving as a superiorsoftware component for cloud computing combined along with integrated parts such as MapReduce. Hadoop, which is an open-source implementation of Google MapReduce, including
a distributed file system, provides to the application programmer the abstraction of the map and the reduce. With Hadoop it is easier for organizations to get a grip on the large volumes of data being generated each day, but at the same time can also create problems related to security, data access, monitoring, high availability and business continuity.Recent progress on classic big data networking technologies, e.g., Hadoop and MapReduce, big data technologies in could computing, big data benchmarking projects, and mobile big data networking.

2. Cloud Computing

In Cloud Computing, the word "Cloud" means "The Internet", so Cloud Computing means a type of computing in which services are delivered through the Internet. The goal of Cloud Computing is to make use of increasing computing power to execute millions of instructions per second. Cloud Computing uses networks of a large group of servers with specialized connections to distribute data processing among the servers. Cloud Computing consists of a front end and back end. The front end user's computer and software required to access the cloud network. Back end consists of various computers, servers and database systems that create the cloud. The user can access applications in the cloud network by connecting to the cloud using the Internet.

Fig 1: Cloud Computing

3. **Hadoop**

Hadoop is an open-source software framework for storing and processing big data in a distributed fashion on large clusters of commodity hardware. Essentially, it accomplishes two tasks: massive data storage and faster processing.

Open-source software:Open source software differs from commercial software due to the broad and open network of developers that create and manage the programs. Traditionally, it's free to download, use and contribute to, though more and more commercial versions of Hadoop are becoming available.

Framework:In this case, it means everything you need to develop and run your software applications is provided – programs, tool sets, connections, etc.

- Distributed:Data is divided and stored across multiple computers, and computations can be run in parallel across multiple connected machines.
- Massive storage:The Hadoop framework can store huge amounts of data by breaking the data into blocks and storing it on clusters of lower-cost commodity hardware.
- Faster processing:Hadoop processes large amounts of data in parallel across clusters of tightly connected low-cost computers for quick results.

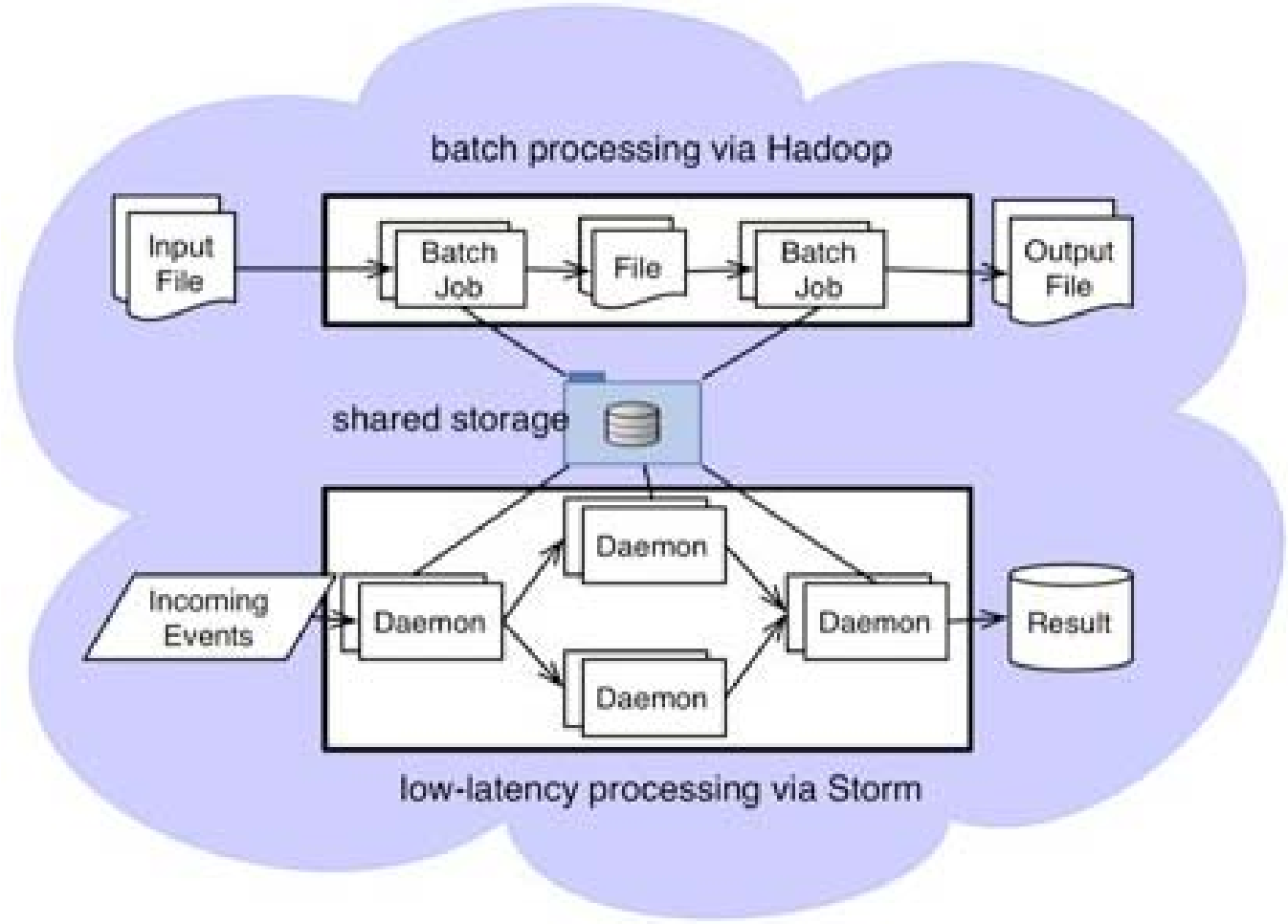

Fig 2: Hadoop

The Hadoop “brand” contains many different tools. Two of them are core parts of Hadoop:

- Hadoop Distributed File System (HDFS) is a virtual file system that looks like any other file system except than when you move a file on HDFS, this file is split into many small files, each of those files is replicated and stored on (usually, may be customized) 3 servers for fault tolerance constraints.
- Hadoop MapReduce is a way to split every request into smaller requests which are sent to many small servers, allowing a truly scalable use of CPU power (describing MapReduce would worth a dedicated post).

Some other components are often installed on Hadoop solutions:

- HBase is inspired from Google’s BigTable. HBase is a non-relational, scalable, and fault-tolerant database that is layered on top of HDFS. HBase is written in Java. Each row is identified by a key and consists of an arbitrary number of columns that can be grouped into column families.
- ZooKeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services. Zookeeper is used by HBase, and can be used by MapReduce programs.

4. Big Data

Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process them using traditional data processing applications.Big data is a buzzword, or catch-phrase, used to describe a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database andsoftware techniques. In most enterprise scenarios the data is too big or it moves too fast or it exceeds current processing capacity.

Fig 3: BigData

Big data has the potential to help companies improve operations and make Big data is difficult to work with using most relational database management systems and desktop statistics and visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers".[13] What is considered "big data" varies depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyze the data set in its domain. Big Data is a moving target; what is considered to be "Big" today will not be so years ahead. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options.

4.1 Big Data Characteristics

Big data can be described by the following characteristics:

- Volume
- Variety
- Velocity
- Variability
- Veracity
- Complexity

Volume – The quantity of data that is generated is very important in this context. It is the size of the data which determines the value and potential of the data under consideration and whether it can actually be considered as Big Data or not. The name 'Big Data' itself contains a term which is related to size and hence the characteristic.

Variety - The next aspect of Big Data is its variety. This means that the category to which Big Data belongs to is also a very essential fact that needs to be known by the data analysts. This helps the people, who are closely analyzing the data and are associated with it, to effectively use the data to their advantage and thus upholding the importance of the Big Data.

Velocity - The term 'velocity' in the context refers to the speed of generation of data or how fast the data is generated and processed to meet the demands and the challenges which lie ahead in the path of growth and development.

Variability - This is a factor which can be a problem for those who analyse the data. This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

Veracity - The quality of the data being captured can vary greatly. Accuracy of analysis depends on the veracity of the source data.

Complexity - Data management can become a very complex process, especially when large volumes of data come from multiple sources. These data need to be linked, connected and correlated in order to be able to grasp the information that is supposed to be conveyed by these data. This situation, is therefore, termed as the 'complexity' of Big Data.

5. Big Data Analytics

Big data analytics is the process of examining large data sets containing a variety of data types i.e.,big data -- to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. The analytical findings can lead to more effective marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over rival organizations and other business benefits.The primary goal of big data analytics is to help companies make more informed business decisions by enabling data scientists, predictive modelers and other analytics professionals to analyze large volumes of transaction data, as well as other forms of data that may be untapped by conventional business intelligence (BI)programs. That could include Web server logs and Internet clickstream data, social media content and social network activity reports, text from customer emails and survey responses, mobile-phone call detail records and machine data captured by sensors connected to the Internet of Things. Some people exclusively associate big data with semi-structured and unstructured data of that sort, but consulting firms like Gartner Inc. and Forrester Research Inc. also consider transactions and other structured data to be valid components of big data analytics applications.

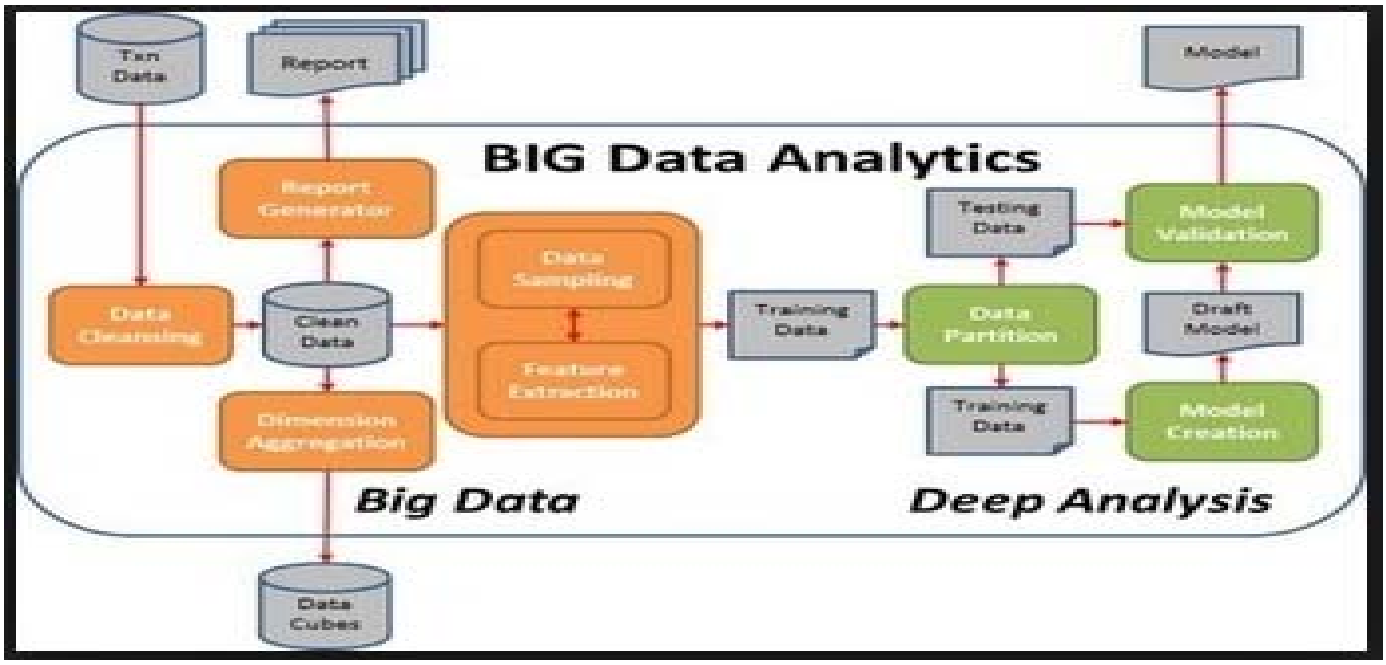

Fig 4: Bigdata analytics

Big data can be analyzed with the software tools commonly used as part of advanced analytics disciplines such as predictive analytics, data mining, text analytics and statistical analysis. Mainstream BI software and data visualization tools can also play a role in the analysis process. But the semi-structured and unstructured data may not fit well in traditional data warehouses based on relational databases. Furthermore, data warehouses may not be able to handle the processing demands posed by sets of big data that need to be updated frequently or even continually for example, real-time data on the performance of mobile applications or of oil and gas pipelines.

6. Big Data Hadoop Architecture

Hadoop was initially inspired by papers published by Google outlining its approach to handling an avalanche of data, and has since become the de facto standard for storing, processing and analyzing hundreds of terabytes, and even petabytes of data.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.

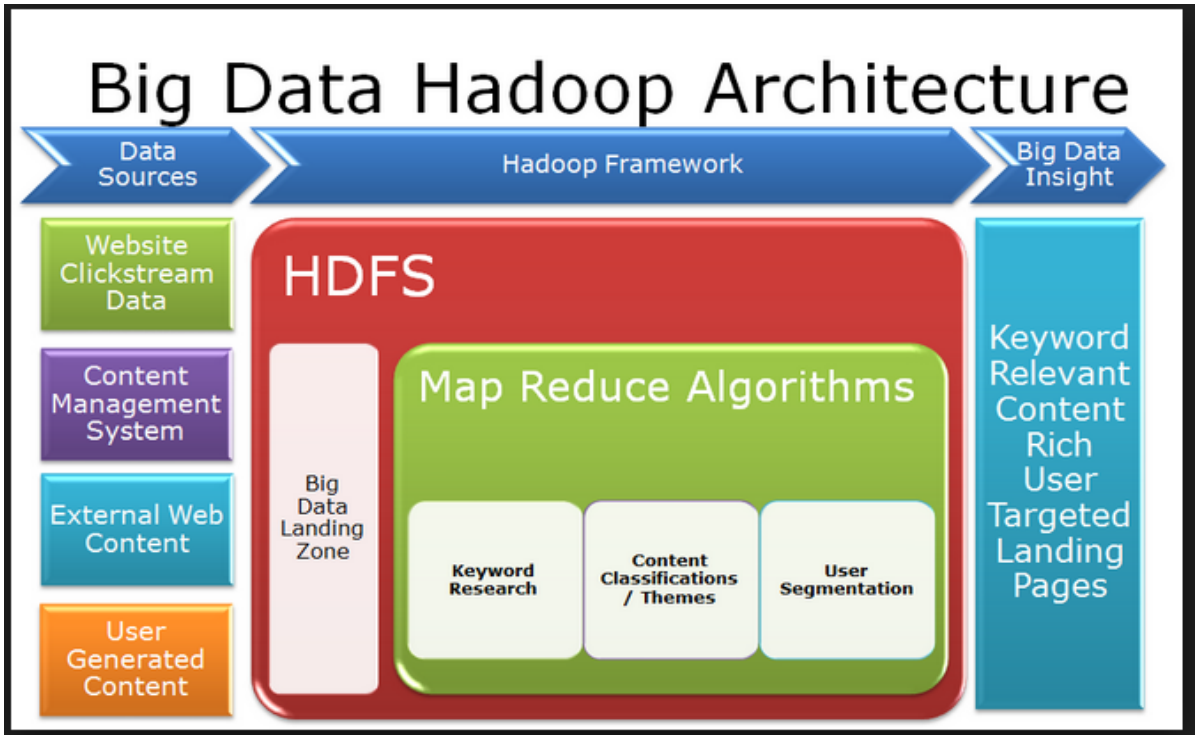

Fig 5: Bigdata Hadoop Architecture

One of the cost advantages of Hadoop is that because it relies in an internally redundant data structure and is deployed on industry standard servers rather than expensive specialized data storage systems, you can afford to store data not previously viable. And we all know that once data is on tape, it's essentially the same as if it had been deleted - accessible only in extreme circumstances. Enterprises who build their Big Data around Cloudera can afford to store literally all the data in their organization, and keep it all online for real-time interactive querying, business intelligence, analysis and visualization.

6.1 The Advantages of Real-Time Big Data Analytics

Big data, the software packages provide a rich set of tools and options where an individual could map the entire data landscape across the company, thus allowing the individual to analyze the threats he/she faces internally.

There are some common characteristics of big data, such as

a) Big data integrates both structured and unstructured data.

b) Addresses speed and scalability, mobility and security, flexibility and stability.

c) In big data the realization time to information is critical to extract value from various data

sources, including mobile devices, radio frequency identification, the web and a growing list of automated sensory technologies. All the organizations and business would benefit from speed, capacity, and scalability of cloud storage. Moreover, end users can visualize the data and companies can find new business opportunities.If big data are combined with predictive analytics, it produces a challenge for many industries. The combination results in the exploration of these four areas:

a) Calculate the risks on large portfolios

b) Detect, prevent, and re-audit financial fraud

c) Improve delinquent collections

d) Execute high value marketing campaigns

7. Security and Issues in Big Data

Cloud computing comes with numerous security issues because it encompasses many technologies including networks, databases, operating systems, virtualization, resource scheduling, transaction management, load balancing, concurrency control and memory management. Hence, security issues of these systems and technologies are applicable to cloud computing. For example,it is very important for the network which interconnects the systems in a cloud to be secure. In addition, resource allocation and memory management algorithms also have to be secure. The big data issues are most acutely felt in certain industries, such as telecoms, web marketing and advertising, retail and financial services, and certain government activities. The data explosion is going to make life difficult in many industries, and the companies will gain considerable advantage which is capable to adapt well and gain the ability to analyze such data explosions over those other companies. Finally, data mining techniques can be used in the malware detection in clouds. The challengesof security in cloud computing environments can be categorized into network level, user authentication level, data level, and genericissues.

Network level: The challenges that can be categorized under a network level deal with network protocols and network security, such as distributed nodes, distributed data, Internode communication.

Authentication level: The challenges that can be categorized under user authentication level deals with encryption/decryption techniques, authentication methods such asadministrative rights for nodes, authentication of applications and nodes, andlogging.

Data level: The challenges that can be categorized under data level deals with data integrity and availability such as data protection and distributed data.

Generic types: The challenges that can be categorized under general level are traditional security tools, anduse of different technologies.

8. Conculsion

Cloud environment is widely used in industry and research aspects; therefore security is an important aspect for organizations running on these cloud environments. Using proposed approaches, cloud environments can be secured for complex business operations.Using big data tools to analyze the massive amount of threat data received daily, and correlating the different components of an attack, allows a security vendor to continuously update their global threat intelligence and equates to improved threat knowledge and insight.Customers benefit through improved, faster, and broader threat protection. By reducing risk,they avoid potential recovery costs, adverse brand impacts, and legal implications.we conclude that promising progresses have been made in the area of big data and big data networking, but much remains to be done. Almost all proposed approaches are evaluated at a limited scale, for which the reported benchmarking projects can act as a helpful compensation for larger-scale evaluations. Moreover, software-oriented studies also need to systematically explore cross-layer, cross-platform tradeoffs and optimizations.

References

1 K, Chitharanjan, and Kala Karun A. "A review on hadoop — HDFS infrastructure extensions.". JeJu Island: 2013, pp. 132-137, 11-12 Apr. 2013.

2 F.C.P, Muhtaroglu, Demir S, Obali M, and Girgin C. "Business model canvas perspective on big data applications." Big Data, 2013 IEEE International Conference, Silicon Valley, CA, Oct 6-9, 2013, pp.32 - 37.

3 Zhao, Yaxiong , and Jie Wu. "Dache: A data aware caching for big-data applications using the MapReduce framework." INFOCOM, 2013 Proceedings IEEE, Turin, Apr 14-19, 2013, pp. 35 - 39.

4 Xu-bin, LI , JIANG Wen-rui, JIANG Yi, ZOU Quan "Hadoop Applications in Bioinformatics." Open Cirrus Summit (OCS), 2012 Seventh, Beijing, Jun 19-20, 2012, pp. 48 - 52.

5 Bertino, Elisa, Silvana Castano, Elena Ferrari, and Marco Mesiti. "Specifying and enforcing access control policies for XML document sources." pp 139-151.

6 E, Bertino, Carminati B, Ferrari E, Gupta A , and Thuraisingham B. "Selective and Authentic ThirdParty Distribution of XML Documents."2004, pp. 1263 - 1278.

7 Kilzer, Ann, Emmett Witchel, Indrajit Roy, Vitaly Shmatikov, and Srinath T.V. Setty. "Airavat:Security and Privacy for MapReduce."

8 Securing Big Data: Security Recommendations for Hadoop and NoSQL Environments."Securosis blog, version 1.0 (2012)

9 P.R , Anisha, Kishor Kumar Reddy C, Srinivasulu Reddy K, and Surender Reddy S. "Third Party Data Protection Applied To Cloud and Xacml Implementation in the Hadoop Environment With Sparql."2012. 39-46, Jul – Aug. 2012.

10 P. Mell and T. Grance, "The NIST Definition of Cloud Computing," 2010.http://www.blogjava.net/zamber/archive.

11 R. L. Grossman, "The Case for Cloud Computing, IT Professional, Vol. 11, No. 2, 2009, pp. 23-27.

12 G. Boss, P. Malladi, D. Quan, L. Legregni and H. Hall,"Cloud Computing," IBM White Paper,2007.http://download.boulder.ibm.com/ibmdl/pub/software/dw/wes/hipods/Cloud_computing_wp_final_8Oct.pdf.

13 Khalid A (2010) Cloud Computing: applying issues in Small Business. International Conference on Signal Acquisition and Processing (ICSAP'10). 278-281.

14 KPMG (2010) From hype to future: KPMG's 2010 Cloud Computing survey..Available: http://www.techrepublic.com/whitepapers/from-hype-to-future-pmgs-2010-cloud-computing-survey/2384291.

15 Rosado DG, Gómez R, Mellado D, Fernández-Medina E (2012) Security analysis in the migration to cloud environments. Future Internet 4(2):469-487.