International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 06 | June-2018 www.irjet.net p-ISSN: 2395-0072

A study on Security and Privacy threats in Big Data with hybrid cloud

Preeti Yadav[1], Poornima Chourasia[2]

[1]Mphil institute of computer science vikram university
[2]Mphil institute of computer science vikram university

Abstract – *In recent years, large-scale growing data have appeared to meet the demands of high storage, supercomputer and application using very large datasets. The occurrence of big data offers the potential for analysis and processing of large datasets. Big data also we can say that belong capacity of data. Today is the need for the new technology for processing large data sets. Apache Hadoop is the good option and it has many components that work together to make Hadoop ecosystem robust and efficient. A hybrid cloud allows different personas to work with data and analytics capabilities where the data and analytics capabilities should be placed in the hybrid cloud environment. In this paper, we aim to analysis on "Big data" Security and Privacy aspects with hybrid cloud. We also provide a review of existing security and privacy protocols for big data. Privacy is one of the critical concerns that hinder the adoption of public cloud. For a simple application, like storage, encryption can be used to protect user's data. But for outsourced data processing, i.e., big data processing with MapReduce framework, there is no satisfying solution. Users have to trust the cloud service providers that they will not leak users' data. This paper refers to Security and Privacy aspects healthcare in big data. The Comparative study between various recent techniques of big data security and privacy approach as well.*

Key Words: Cloud computing, Hybrid cloud, Big data, Security and Privacy perspective.

1. INTRODUCTION

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction (Mell & Grance, 2009). These feature of cloud computing make it attractive for many applications, e.g., database system, customer relationship management, and call centre. Especially, as an emerging technology, big data attracts attention from both the industrial and academic communities.

Security has been considered as one of the critical concerns that hinder the wide adoption of public cloud, especially for the enterprises and the government market. In reality, it is very unlikely that the companies like Amazon, Microsoft, and Google who run the cloud service will try to access the users' data without permission. The main threats come from malicious users and administrators. Due to the virtualization technology used in cloud computing, malicious users may cross the boundary to access others' data (Ristenpart, Tromer, Shacham & Savage, 2009). Administrators usually have higher privileges and they may abuse this ability to learn users' data without permission (like the Snowden case (Toxen, 2014)). For simple cloud service like data storage, the user can protect the data from these threats with encryption. More complex techniques are designed to support applications like sharing and efficient retrieval (Thuraisingham, Khadilkar, Gupta, Kantarcioglu & Khan, 2010). Applications that are dependent on both the storage and computation capability of cloud need a more complex solution for the security concern. Cloud computing is a successful model of service focused on computing and modernized the way computing infrastructure is abstracted The three most popular cloud services includes is Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). The impression can also be stretched to the Database as a Service as a Service as a Service. Flexibility, pay-per-utilize, low forthright venture, ease to market, and exchange of dangers are a portion of the major empowering highlights that make distributed computing a pervasive worldview for conveying novel applications that were not financially possible in conventional endeavour foundation settings. Accessible database management systems (DBMS) both for update rigorous application workloads, as well as decision support systems are a critical part of the cloud infrastructure. Cloud can be classified into three type are public cloud, private cloud, hybrid cloud.

1.1 Security issued in cloud computing

The main factor for IT Executives when it moves to cloud computing is security and privacy. Its environments are the multi-domain environment in which various resources are shared. While sharing Hardware and placing data it seems to be a high risk factor. Any unauthorized person can easily be hacked either accidentally or due to malevolent attack. Hence data storage would be a major security violation. Cloud adoption is accelerating rapidly, driven by the cost savings, agility, and other advantages it offers. As you transition to the cloud, the shared security responsibility model means that you must secure what you put IN the cloud, and that your security solution meets internal and regulatory compliance rules. Security is optimized for leading cloud service providers (CSPs) including AWS, Microsoft Azure, Google Cloud, and more. It makes using leading orchestration tools like Chef, Puppet,

SaltStack, Ansible, and AWS Opworks easy, providing deployment examples and automated generation of policy scripts that enable security to be managed as part of cloud operations.

1.2 Hybrid cloud

When you wish to maintain different business applications with different levels of security particularly this service is useful. Hybrid clouds services are a combination of public and private clouds implemented by different providers. One of the disadvantages of these services is that we have to manage different security platforms together. Where some data resides in the private cloud environment and some resides in the public cloud environment. Here, public clouds are used for tasks that are not so sensitive while private clouds are used for those that are vital. It is more like a middle path and apparently more suitable for companies who are just making the cloud computing plunge.

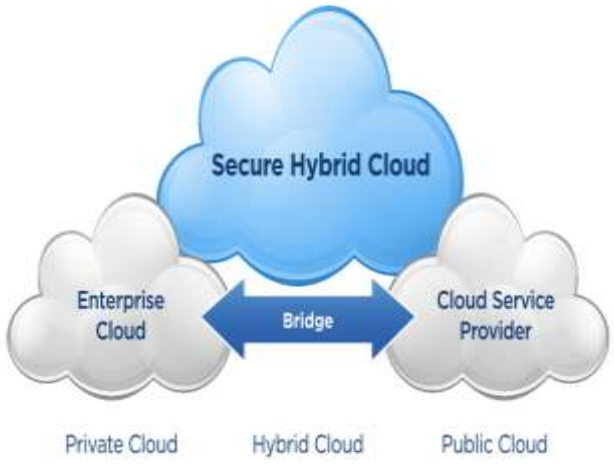

Fig1: Hybrid cloud

When it comes to security, there are indeed some niggling concerns associated with cloud hosting since it is a domain that resides in a common location. However, it is surely emerging as a highly favorable concept and futuristic too. The hybrid infrastructure offers the best of both worlds, private and public.

The cloud infrastructure is indeed a huge boon to the computing world.It is usually a combination of on-and off premise. A comparison of the different issues of cloud computing on cloud deployment models is given below.

Table 1 .Cloud Deployment models and issues

Model	Security Issues	Cost issues	Control issues	Legal Issues
Public Cloud	Least Secure, Multi-tenancy, Transfers over the net	Setup needs highest, usage is lowest pay for what we use	Least control	Jurisdiction of storage
Private Cloud	Most Secure	Setup need high, new operational processer are required	Most control	-
Hybrid Cloud	Control of Security between Public and Private cloud	-	Least Control	Jurisdiction of storage

From the above table it is understood that private cloud has more secure than public. According to the cost issues the setup cost is very high in both private and public. If we consider control issues private cloud is most control than public and hybrid clouds. But public and hybrid clouds have same legal issues.

1.3 Big data

Big Data computing is an emerging data science paradigm of multidimensional information mining for scientific discovery and business analytics over large-scale infrastructure. The data collected or produced from several scientific explorations and business transactions often require tools for effective data management, analysis, validation, visualization, and dissemination, preserving the intrinsic value of the data. Before “Big data is high-volume, high-velocity and high-variety, information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making”.

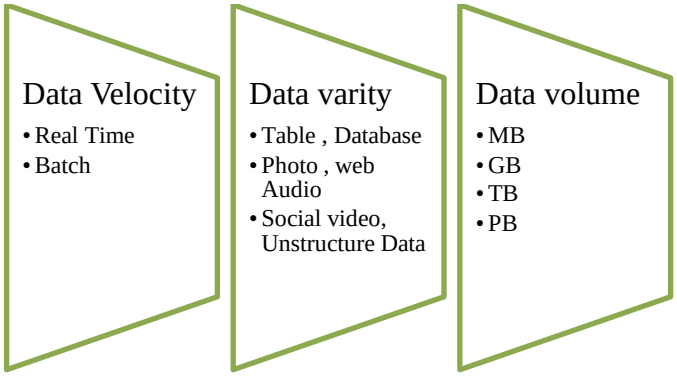

Fig2: Big data three v's

Big Data and Cloud Computing are two important staples in the contemporary years, empowering computing resources to be provided information technology services with high competence and effectiveness. Now a day's big data is one of the most problems that researchers try to solve it and centering their researches over it to get the problem of how big data could be supervision in the recent systems and managed with the cloud of computing, and one of the most important issues is how to gain a perfect security for large data in cloud computing, cloud computing security and the mechanisms that used to protect and secure data in large data with an available clouds. As now a day data increased from day to day the most problem of data is storage and data security.

As big data has some key data security issues as they are distributed frameworks, non-relational data store, storage, endpoint, real-time security/compliance tools, data mining, access controls, granular auditing and data origin. Big data open source tools are Hadoop.

Hadoop is an open source big data and a Java-based programming framework that are supports data communication in a distributed computing environment. As Hadoop make in Apache Software Foundation Hadoop makes it possible to run applications on frameworks with a huge number of hubs or group including a large number of terabytes. The present day Apache Hadoop innovation includes of the Hadoop bit, MapReduce, the Hadoop distributed file system (HDFS) and a number of related ventures form, for example, Apache Hive, HBase and Zookeeper and so forth. The Hadoop structure significantly is utilized by real organizations including Google, Yahoo, and IBM as they favored working frameworks are Windows 8, 10, Linux and Macintosh.

MapReduce is an in indispensable and important element part of the Apache Hadoop software framework which divides data into subpart. Hadoop agrees the robust and distributed allowance of the massive unstructured data sets across the commodity computer clusters in through each node of the cluster includes its own storage. MapReduce to serves have two critical into tasks as it correspondences out work to various nodes within the cluster or map, it systematizes and reduces the results from each node to an uncertain answer to a query.

2. Security and Privacy Challenges in Big data using MapReduce

Security and privacy techniques for processing big-data have to deal with huge amount of data, possibly arriving at high speed from different sources. Moreover, in MapReduce computations, data is partitioned into small-sized splits that are replicated and distributed to several nodes. Each split has to be transferred in a secure and private manner. This replicated and distributed nature constitutes unique challenges in terms of data storage security, as compared to a system that holds the whole data in a single place. Challenges are facing are Highly distributed nature of MapReduce computations, Data flow, Scalability, fault tolerance, and transparency, Economical issues and entrusted data access. All the above challenges to MapReduce framework in clouds indicate new security and privacy requirements.

2.1 Security Aspects in MapReduce

The security of data and computations plays a significant role in MapReduce computations on both hybrid and public clouds. Without security, MapReduce computations as well as MapReduce infrastructures can be affected by several types of attacks. In this section, we present security threats and security requirements for MapReduce computations. Notice that even though some security threats and security requirements are common for MapReduce and for generic cloud computing, we will focus on security threats and security requirements in the context of MapReduce. Security threats that can harm a MapReduce computation and the framework in the absence of secure MapReduce environment. Distributed and replicated data processing in MapReduce open an opportunity for a wide range of attacks. While those attacks follow the same ideas as attacks in different cloud computation models, the exact application is different for MapReduce attack are Impersonation attack, Denial-of-Service (DoS) attack, Replay attack, Eavesdropping, Man-in-the-Middle (MiM) attacks and Repudiation.

2.2 Privacy Aspects in MapReduce

Privacy ensures that sensitive data is not exposed to untrusted users and trespassers (i.e., cloud providers, other data providers, users of MapReduce, or adversaries). Notice that the data providers are interested in allowing some sorts of computations on the data, however, there is also a requirement to preserve breach of sensitive data. Sensitive data in this case is case specific and might be personal records with identifier information (personally identifiable information PII), organization specific information and etc. computing is an emerging data science paradigm of multi-dimensional information mining for scientific discovery and business analytics over large scale infrastructure. The data collected or produced from several scientific explorations and business transactions often

require tools for effective data management, analysis, validation, visualization and dissemination, preserving the intrinsic value of the data. Cloud computing and the deployment of MapReduce on public clouds present a new set of challenges the privacy of data. Here, we describe privacy challenges of cloud computing in the context of MapReduce and divide them into a few cases according to adversarial behaviors of public clouds and users.

3. Proposed Solution for security and Privacy MapReduce

The security of data and computations play a significant role in MapReduce computations on hybrid cloud. ClusterBFT uses the Byzantine Failure Tolerant (BFT) replication technique to cope with a situation where the cloud is trusted but there are potentially malicious nodes or users in a cluster. BFT replication is used for computational results verification and for overcoming untrusted, possibly malicious nodes. BFT replication techniques perform calculations in parallel on multiple replicas, then compare all the produced outputs to identify erratic behavioral nodes and decide a correct output based on a majority vote. However, current BFT replication techniques were developed for stand-alone servers and do not suit cloud-based computations, where data flow among different nodes and a computation consists of a number of stages to be performed on different nodes, as it is done in MapReduce. In order to overcome this gap, ClusterBFT algorithm adopts BFT replication for highly-scalable, distributed and high-granularity cloud computations.

3.1 Proposed Solutions for Privacy in MapReduce

Information privacy is the privilege to have some control over how personal information is collected and used. When data are stored on cloud, if data confidentiality or integrity is breached it will have a direct effect on user privacy. Some framework of MapReduce privilege privacy is as follows HybrEx, Sedic, Tagged-Mapreduce, Semrod, Promrtheus.

HybrEx: Hybrid Execution (HybrEx)[22] is the first MapReduce framework designed for the hybrid cloud. In HybrEx, data is divided into sensitive and non-sensitive data, non-sensitive data is sent to public clouds while sensitive data is kept in a private cloud.

HybrEx allows four types of execution models of MapReduce computations, as follows:

•Map hybrid: the map stage is executed at both public and private clouds, notwithstanding, the decrease stage is executed at a private cloud as it were.

•Horizontal partitioning: the guide organize is executed (on encrypted data) at public clouds just, while the reduce arrange is executed at a private cloud.

•Vertical partitioning: the map phase and the reduce phase are executed on both public and private clouds while data transmission between private and public clouds is not allowed.

•Hybrid: the guide arrange and the reduce organize are executed on both public and private clouds and data transmission among clouds is furthermore possible.

Two dependability check models, to be particular full uprightness checking and smart trustworthiness checking, are in like manner suggested. Not with standing, HybridEx does not deal with a key that is made at public and private clouds in the guide organize.

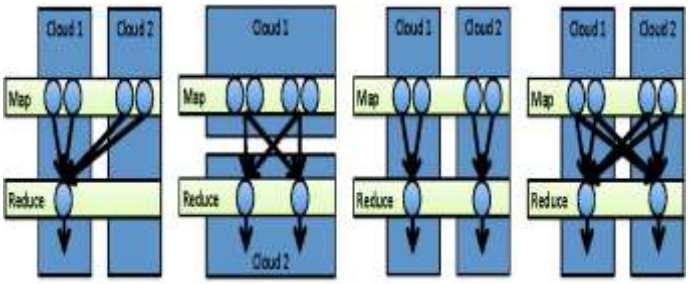

Fig. 2: Execution Categories for HybrEx MapReduce

Sedic. In order to solve key problem of HybridEx, Sedic [129] provides strategic data movement from the map phase that executes on public clouds to the reduce phase that executes on a private cloud, by using an automatic analysis and transformation of the reduce code. In order to decrease the communication between a public cloud and a private cloud, outputs of the map phase (at the public cloud) are aggregated before their transmission to the private cloud. In addition, Sedic framework automatically partitions a job by following security levels of data and distributes a job between private and public clouds.

Tagged-MapReduce: HybrEx consider data affectability before an occupation's execution. Tagged-MapReduce[23] recognizes data-affectability in the midst of execution of a job, where the guide organize and the diminishing stage are executed on public and private clouds. The framework handles sensitivity of intermediate outputs that may contain sensitive data, and hence, cannot be processed by the reduce phase at public clouds. Two policies, non-upgrading policy and downgrading policy, help in identifying on-the-fly data sensitivity, and four scheduling modes (single-phase, two-phase crossing, two-phase non-crossing, and hand-off modes), assign outputs of the map phase to reducers regarding data sensitivity. In addition, Tagged-MapReduce supports iterative MapReduce jobs. However, HybrEx ,Sedic , and Tagged-MapReduce are unable to handle the situation efficiently when a key is generated at public and private clouds. In order to solve this, Oktay et al. suggested Secure and Efficient MapReduce Over hybriD clouds (SEMROD) that prevents the leakage of sensitive data and efficiently exploits public

resources for executing a given single (or multi-level) MapReduce job.

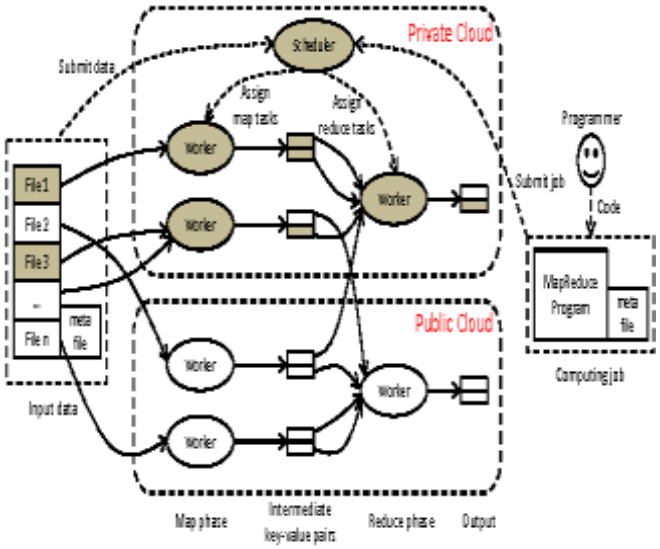

Fig4: Tagged-MapReduce

SEMROD. SEMROD first finds sensitive and non-sensitive data and sends non-sensitive data to public clouds. Private and public clouds execute the map phase. However, instead of sending only outputs of the map phase containing sensitive keys to the private cloud, the private cloud pulls all the outputs, but executes the reduce phase operation only on record associated with sensitive keys and ignores non-sensitive keys. Public clouds execute the reduce phase on all the outputs. Hence, they are unable to know the sensitive keys. In the end, a filtering step removes duplicate entries, creating by sensitive key.

Prometheus. In order to outsource non-sensitive data, which is stored in relations, to public clouds, Prometheus removes quasi-identifiers (a quasi-identifier refers to a subset of attributes that can uniquely identify most tuples in a relation) using a hypergraph. After the discovery of quasi-identifiers, attributes are distributed over public clouds, and an attribute location table is used to store the name of relations and the location of relations-attributes. This allows the system to ensure that no sensitive data is stored in untrusted public clouds. It also avoids heavy workload on reducers at the user-end by sending merged outputs of public clouds and a mapping table of tuples from the private cloud to the user. Reducers (at the user-end) construct the final output. On the downside, Prometheus allows only search operations on a hybrid cloud.

3. CONCLUSIONS

Processing a huge amount of data is not simple using the classical parallel computing, due to the failure of computing nodes and scalability of the system. MapReduce has the gives an effective, blame tolerant, versatile, and straightforward preparing of huge scale data. However, MapReduce was not designed to be deployed on hybrid clouds, where security and privacy of data and computations are two prime concerns. Since public clouds provide an easy way for computations and storage, a number of algorithms and frameworks regarding security and privacy of data-computations were developed for executing a MapReduce job on hybrid clouds. In this survey, we have got to now security and privacy challenges and requirements in MapReduce. Security attacks in MapReduce impersonation, denial-of-services, replay, eavesdropping, man-in-the-middle, and repudiation attacks are presented. show how they can impact a MapReduce computation. We believe that in the future we will have MapReduce frameworks that provide multiple types of computations in information secure manner.

REFERENCES

[1] "S.VikramPhaneendra , E.Madhusudhana Reddy "Big Data - Solutions for RDBMS Problems - A Survey " , In twelfth IEEE/IFIP Network Operations and Management Symposium (NOMS 2010) (Osaka, Japan, Apr 19{23 2013) distributed on 19 Apr 2013."

[2] " Kiran kumara Reddi and Dnvsl Indira "Diverse Technique to Transfer Big Data : overview " , IEEE Transactions on 52(8) (Aug.2013) 2348 { 2355} India vol.8 distributed on aug 2013."

[3] "Kilzer , Ann , Emmett Witchel, Indrajit Roy , VitalyShmatikov and Srinath T.V. Setty. "Airaval: security and Privacy for MapReduce " NSDI April 28 2010 ,San jose ,CA ,USA."

[4] "MrigankMridul, AkashdeepKhajuria, SnehasishDutta, Kumar N " Analysis of Bigdata using Apache Hadoop and Map Reduce" Volume 4, Issue 5, 27 May 2014"."

[5] "Yousef K. Sinjilawi, Mohammad Q. AL-Nabhan and Emad A. Abu-Shanab "Addressing Security and Privacy Issues in Cloud Computing " , Journal of Emerging Technologies in Web Intelligence, Vol. 6, No. 2, published on May 2014."

[6] "Tapan P. Gondaliya, Dr. Hiren D. Joshi "Big Data difficulties and Hadoop as one of the arrangement of big data with its Modules", International Journal of Scientific and Engineering Research,Volume5,Issue6,June-2014 ISSN 2229-5518 ."

[7] "Balachandar.R , Tharini.N, FashilaParveen.S "A survey on the value of big data-The next big thing in Information" , International Journal of Advanced Engineering and Recent Technology Volume 3 Issue 1, January 2016 ."

[8] "M Shilpa and ManjitKaur "BIG Data and Methodology-An audit " , International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 10, October 2013."

[9] "Venkatesh H, Shrivatsa D Perur, NiveditaJalihal "A Study on Use of Big Data in Cloud Computing Environment " , IJCSIT International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 2076-2078."

[10] "Pedro CaldeiraNeves, Jorge Bernardino "Big Data in the Cloud: A Survey" , OJBD ISSN 2365-029X, Volume 1, Issue 2, 2015."

[11]"http://blog.cloudera.com/blog/2011/09/snappy-and-hadoop"

[12]"Apache Ranger, available at: http://ranger.incubator.apache.org. "

[13]" W. Wei, J. Du, T. Yu, and X. Gu. SecureMR: A service integrity assurance framework for MapReduce. In Twenty-Fifth Annual Computer Security Applications Conference, ACSAC 2009,Honolulu, Hawaii, 7-11 December 2009, pages 73–82, 2009."

[14]" M. R. Randazzo, M. Keeney, E. Kowalski, D. Cappelli, and A. Moore. Insider threat study: Illicit cyber activity in the banking and finance sector, 2005"

[15]"D. Das, O. O'Malley, S. Radia, and K. Zhang. Adding security to Apache Hadoop. HortonworksTechnical Report 1, 2010"

[16]" K. Zhang, X. Zhou, Y. Chen, X. Wang, and Y. Ruan. Sedic: privacy-aware data intensive computing on hybrid clouds. In Proceedings of the 18th ACM Conference on Computer and Communications Security, CCS 2011."

[17]" M. Pastore, M. Pastore, and E. Dulaney. CompTIA Security+ Study Guide: Exam SY0-101 Wiley, 2006."

[18]"Apache sentry available at http://sentry.incubator.apache.org"

[19]" Q.Shen, L.Zhang, X.Yang, Y.Yang, Z.Wu, and Y.Zhang. SecDM: Securing data migration between cloud storage systems. In IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing, DASC 2011, 12-14 December 2011, Sydney, Australia."

[20]" Q. Shen, Y.Yang, Z.Wu, X.Yang, L.Zhang, X.Yu, Z.Lao, D.Wang, and M. Long. SAPSC: security architecture of private storage cloud based on HDFS. In 26th International Conference on Advanced Information Networking and Applications Workshops, WAINA 2012, Fukuoka, Japan,March 26-29, 2012."

[21]" J.J.Stephen and P.Eugster. Assured cloud-based data analysis with ClusterBFT. In Middleware 2013 - ACM/IFIP/USENIX 14th International Middleware Conference, Beijing, China, December 9-13, 2013."

[22]" S. Y. Ko, K. Jeon, and R. Morales. The HybrEx model for confidentiality and privacy in cloud computing. In 3rd USENIX Workshop on Hot Topics in Cloud Computing, HotCloud'11, Portland, OR, USA, June 14-15, 2011."

[23]" C. Zhang, E. Chang, and R. H. C. Yap. Tagged-MapReduce: A general framework for secure computing with mixed-sensitivity data on hybrid clouds. In 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, CCGrid 2014, Chicago, IL, USA, May 26-29,2014"

[24]" E. Blass, R. D. Pietro, R. Molva, and M. Önen. PRISM - privacy-preserving search in MapReduce. In Privacy Enhancing Technologies - 12th International Symposium, PETS 2012, Vigo, Spain, July 11-13, 2012."

[25]" E. Blass, G. Noubir, and T. V. Huu. EPiC: Efficient privacy-preserving counting for MapReduce,2012."

[26]" R. A. Popa, C. M. S. Redfield, N. Zeldovich, and H. Balakrishnan.CryptDB: processing queries on an encrypted database. Commun. ACM, 55(9):103–111, 2012."