Big Data in the Cloud Environment and Enhancing Model for data security using Encrypted Cloud Servers

S.Praveen Kumar, Dr.Y.Srinivas, Dr Vamsi Krishna, Ashish Kumar, Harsha Vardhan, T. Dhananjaya Rao

***Abstract*— Cloud computing has been a revolutionary advancement towards the technological development of information technology. Further, big data took the benefits of cloud services in order to provide better storing, handling and extracting data of different kinds (structured, semi-structured and unstructured). Apache spark is a lightning fast cluster which has enhanced the computing capabilities of Hadoop by providing faster results. In this paper we also discuss about the encrypted cloud data and propose a secure and privacy preserving algorithm for the same.**

Encrypted cloud data contains a secure mechanism to secure the actions happening in the cloud server. The Data Owner uploads the data into the cloud, whereas the Data User searches for the required file and then asks for the permission to download it from the Data Owner. The Data Owner and Data User both have their own secret key which helps them to authenticate themselves.

***Index Terms*— Cloud, Big Data, Hadoop, Spark, Encrypted cloud data.**

I. INTRODUCTION

The rise of cloud computing and cloud data stores has been a precursor and facilitator to the emergence of big data. [1]Cloud computing is the commodification of computing time and data storage by means of standardized technologies. Before the cloud era, these tasks were expensive, technically challenging and possible to only a few.[2] Cloud computing refers to the use of a scalable public computer network to perform computing tasks. E.g., Amazon Web services, Azure, Google Cloud.[3]Whereas big data deals with how to organize those data which are abundant in size and label them into different kinds such as structured, semi-structured or unstructured. Big data also includes the technologies involved in performing these tasks. It is a new paradigm of collecting, storing, handling and extracting meaning from different kinds of data.[4]Big data is the science of analysing high volumes of diverse data in near-real time (volume, velocity, variety). Typically it involves using NoSQL technology and a distributed architecture to analyse the data.[5] The analysis can be done in the public cloud or on private infrastructure.

Cloud computing employs visualization of computing resources to run numerous standardized virtual servers on the same physical machine.[6] Cloud providers achieve with this economies of scale, which permit low prices and billing based on small time intervals, e.g. hourly.[7]

This standardization makes it an elastic and highly available option for computing needs. The availability is not obtained by spending resources to guarantee reliability of a single instance but by their inter-changeability[8] and a limitless pool of replacements. This impacts design decisions and requires dealing with instance failure gracefully.

Cloud Big Data Challenges:

Vertical scaling achieves elasticity by adding additional instances with each of them serving a part of the demand. Software like Hadoop are specifically designed as distributed systems to take advantage of vertical scaling. They process small independent tasks in massive parallel scale.[9] Distributed systems can also serve as data stores like NoSQL databases, e.g. Cassandra or HBase, or filesystems like Hadoop's HDFS. Alternatives like Storm provide coordinated stream data processes in near real-time through a cluster of machines with complex workflows.

The inter-changeability of the resources together with distributed software design absorbs failure and equivalently scaling of virtual computing instances unperturbed. Spiking or bursting demands can be accommodated just as well as personalities or continued growth. Renting practically unlimited resources for short periods allows one-off or periodical projects at a modest expense. Data mining and web crawling are great examples.[10] It is conceivable to crawl huge web sites with millions of pages in days or hours for a few hundred dollars or less. Inexpensive tiny virtual instances with minimal CPU resources are ideal for this purpose since the majority of crawling the web is spent waiting for IO resources. Instantiating thousands of these machines to achieve millions of requests per day is easy and often costs less than a fraction of a cent per instance hour.

Of course, such mining operations should be mindful of the resources of the web sites or application interfaces they mine, respect their terms, and not impede their service.[11] A poorly planned data mining operation is equivalent to a denial of service attack. Lastly, cloud computing is naturally a good fit for storing and processing the big data accumulated form such operations.

II. HADOOP IN CLOUD:

With its unlimited scale and on-demand access to compute and storage capacity, cloud computing is the perfect match for big data processing.[12] Hadoop is an open-source Cloud computing environment that implements the Google MapReduce framework in Java. Hadoop is created and maintained by the Apache project. MapReduce makes it very easy to process and generate large data sets on the cloud. Using MapReduce, you can divide the work to be performed in to smaller chunks, where multiple chunks can be processed concurrently.[13] You can then combine the results to obtain the final result. MapReduce enables one to exploit the massive parallelism provided by the cloud and provides a simple interface to a very complex and distributed computing infrastructure. If you can model your problem as a MapReduce problem, then you can take advantage of the Cloud computing environment provided by Hadoop.[14] Hadoop enables the development of reliable, scalable, efficient, economical and distributed computing using very simple Java interfaces - massive parallel code without the pain. Hadoop includes a distributed file system, HDFS and a system for provisioning virtual Hadoop clusters over a large physical cluster called Hadoop On Demand (HOD).

2.1 Hadoop architecture:

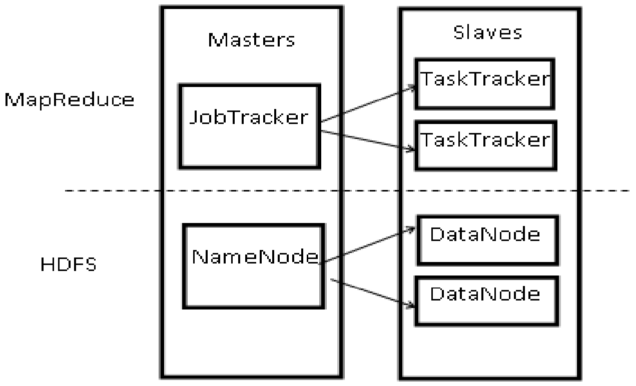


Fig. 2.1. Hadoop Architecture

2.2 Spark and Hadoop belong together:

As data science has matured over the past few years, so has the need for a different approach to data and its “bigness.” There are business applications where Hadoop outperforms the newcomer Spark, but Spark has its place in the big data space because of its speed and its ease of use.[15] This analysis examines a common set of attributes for each platform including performance, fault tolerance, cost, ease of use, data processing, compatibility, and security.[16]

The most important thing to remember about Hadoop and Spark is that their use is not an either-or scenario because they are not mutually exclusive. Nor is one necessarily a drop-in replacement for the other.[17] The two are compatible with each other and that makes their pairing an extremely powerful solution for a variety of big data applications.

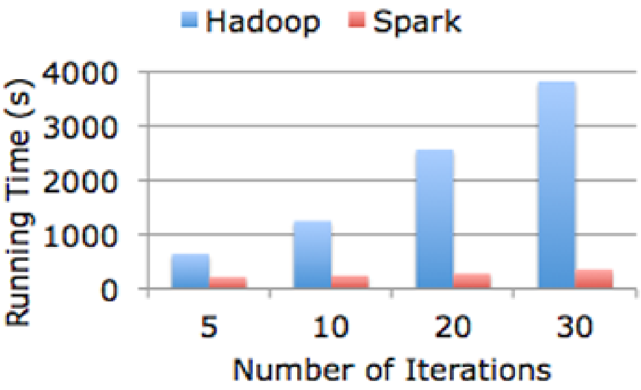


Fig. 2.2 Comparison between Hadoop and Spark runtime.

In the above graph we see that the Hadoop cluster takes a longer running time compared to spark for several numbers of iterations. Spark is 100 times faster than Hadoop in processing and computing data.

III. CLOUD COMPUTING SECURITY:

Cloud computing security or cloud security refers to a broad set of policies, technologies, and controls deployed to protect data, applications, and the associated infrastructure of cloud computing. It is a sub-domain of computer security, network security, and, more broadly, information security.[19]

In order to conserve resources, cut costs, and maintain efficiency, Cloud Service Providers often store more than one customer's data on the same server. As a result, there is a chance that one user's private data can be viewed by other users (possibly even competitors).[18] To handle such sensitive situations, cloud service providers should ensure proper data isolation and logical storage segregation.

Cloud encryption:

Cloud encryption is the transformation of a cloud service customer's data into ciphertext. Cloud encryption is almost identical to in-house encryption with one important difference -- the cloud customer must take time to learn about the provider's policies and procedures for encryption and encryption key management. The cloud encryption capabilities of the service provider need to match the level of sensitivity of the data being hosted.[20]

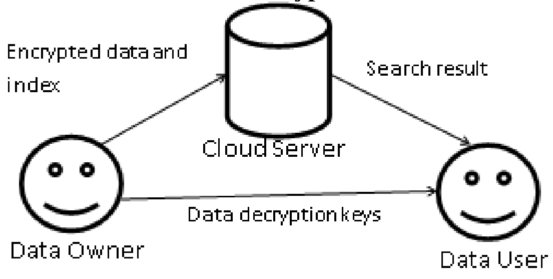
Because encryption consumes more processor overhead, many cloud providers will only offer basic encryption on a few database fields, such as passwords and account numbers.[21] At this point in time, having the provider encrypt a customer's entire database can become so expensive that it may make more sense to store the data in-house or encrypt the data before sending it to the cloud.[22] To keep costs low, some cloud providers have been offering alternatives to encryption that don’t require as much processing power. [23]
Cloud application users have choices when it comes to the strength of their encryption solutions, and standards have emerged across jurisdictions and industries to provide consistency and a level of assurance.[24] Many commercial businesses now follow it because of its maturity and strong level of encryption.

In the past, many businesses felt comfortable allowing the cloud provider to manage encryption keys, believing that security risks could be managed through contracts, controls and audits.[25

IV. .STRUCTURE OF ENCRYPTED CLOUD DATA:

Cloud computing is the long dreamed vision of computing as a utility, where cloud customers can remotely store their data into the cloud so as to enjoy the on-demand high-quality applications and services from a shared pool of configurable computing resources [2], [3]. Its great flexibility and economic savings are motivating both individuals and enterprises to outsource their local complex data management system into the cloud. To protect data privacy and combat unsolicited accesses in the cloud and beyond, sensitive data, for example, e-mails, personal health records, photo albums, tax documents, financial transactions, and so on, may have to be encrypted by data owners before outsourcing to the commercial public cloud [4]; this, however, obsoletes the traditional data utilization service based on plaintext keyword search.[5] Thus, exploring privacy preserving and effective search service over encrypted cloud data is of paramount importance. Considering the potentially large number of on-demand data users and huge amount of outsourced data documents in the cloud, this problem is particularly challenging as it isextremely difficult. [6]

4.1 Architecture of an encrypted cloud



4.1 Architecture of an encrypted cloud

Data owners are a collection of documents and files which are uploaded or modified in the cloud server.
Data users are the authorized persons to access the data uploaded by Data owners into the cloud.
Cloud server stores the data uploaded by the Data owner and computes the search request given by the Data user to give the desired result.

4.2 Proposed algorithm for better encrypted security in cloud environment:

Let us consider that there are ‘n’ data owners and ‘m’ data users. The data owners are responsible for uploading the data into the cloud specifying their category, whereas the data users search for the required data and select the best ranked one. We provide an algorithm for following the whole process:
Step 1: Data user and Data owner registers into the cloud. They both are provided with a secret key which has to be used everytime they login in order to authenticate the actions.
Step 2: The Data owner uploads the data into the cloud specifying its category. This helps the user to provide category wise search.
Step 3: The Data user searches for the required data and if found requests the data owner to approve the download. The search result is given in a ranked fashion which is decided by the number of times the file has been downloaded.
Step 4: the Data owner accepts or rejects the request of the user.
*Step 5:*If the request is approved then the Data user can download the data.
The steps mentioned above help in providing secure and privacy preserving access to the data by the data users and ask permission of download from the data owner. The user searches for the data in the required category and receives the result in a ranked fashion in descending order based on the number of times it has been downloaded. This helps the user to provide an advanced search and helps them to get related files based on the keyword of the search.

V. CONCLUSION:

In this paper, we have discussed thoroughly about the cloud environment and big data influence on in. Big data and Hadoop have just been using the services of cloud in the optimal way and provided a better way of storing, accessing and computing data of any kind (structured, semi-structured, unstructured). Though Hadoop was an advanced and fast way to perform operations on data of huge amount, Apache Spark took over the MapReduce as it was 100 times faster than MapReduce. We used a graph to explain the computational time difference between the both. Cloud computing security has always been a concern since cloud services are widely used and thus result in vary fraud and security breaching. To avoid this we have discussed about the architecture of the encrypted cloud data. Further we proposed an algorithm which explains the working of the structure of encrypted cloud data. This will help to secure the privacy of data uploaded by the Data owner. In our future works we would be working on various algorithms which can provide better security in cloud environment.

REFERENCES:

[1] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, “Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data,” Proc. IEEE INFOCOM, pp. 829-837, Apr, 2011.

[2] L.M. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner, “A Break in the Clouds: Towards a Cloud Definition,” ACM SIGCOMM Comput. Commun.Rev., vol. 39, no. 1, pp. 50-55, 2009.

[3] N. Cao, S. Yu, Z. Yang, W. Lou, and Y. Hou, “LT Codes-Based Secure and Reliable Cloud Storage Service,” Proc. IEEE INFOCOM, pp. 693-701, 2012.

[4] S. Kamara and K. Lauter, “Cryptographic Cloud Storage,” Proc. 14th Int’l Conf. Financial Cryptograpy and Data Security, Jan. 2010.

[5] A. Singhal, “Modern Information Retrieval: A Brief Overview,” IEEE Data Eng. Bull., vol. 24, no. 4, pp. 35-43, Mar. 2001.

[6] I.H. Witten, A. Moffat, and T.C. Bell, Managing Gigabytes: Compressing and Indexing Documents and Images. Morgan Kaufmann Publishing, May 1999.

[7] D. Song, D. Wagner, and A. Perrig, “Practical Techniques for Searches on Encrypted Data,” Proc. IEEE Symp. Security and Privacy, 2000.

[8] E.-J. Goh, “Secure Indexes,” Cryptology ePrint Archive, http:// eprint.iacr.org/2003/216. 2003.

[9] Y.-C. Chang and M. Mitzenmacher, “Privacy Preserving Keyword Searches on Remote Encrypted Data,” Proc. Third Int’l Conf. Applied Cryptography and Network Security, 2005.

[10] R. Curtmola, J.A. Garay, S. Kamara, and R. Ostrovsky, “Searchable Symmetric Encryption: Improved Definitions and Efficient Constructions,” Proc. 13th ACM Conf. Computer and Comm. Security (CCS ’06), 2006.

[11] D. Boneh, G.D. Crescenzo, R. Ostrovsky, and G. Persiano, “Public Key Encryption with Keyword Search,” Proc. Int’l Conf. Theory and Applications of Cryptographic Techniques (EUROCRYPT), 2004.

[12] M. Bellare, A. Boldyreva, and A. ONeill, “Deterministic and Efficiently Searchable Encryption,” Proc. 27th Ann. Int’l Cryptology Conf. Advances in Cryptology (CRYPTO ’07), 2007.

[13] M. Abdalla, M. Bellare, D. Catalano, E. Kiltz, T. Kohno, T. Lange, J. Malone-Lee, G. Neven, P. Paillier, and H. Shi, “Searchable Encryption Revisited: Consistency Properties, Relation to Anonymous Ibe, and Extensions,” J. Cryptology, vol. 21, no. 3, pp. 350- 391, 2008.

[14] J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou, “Fuzzy Keyword Search Over Encrypted Data in Cloud Computing,” Proc. IEEE INFOCOM, Mar. 2010.

[15] D. Boneh, E. Kushilevitz, R. Ostrovsky, and W.E.S. III, “Public Key Encryption That Allows PIR Queries,” Proc. 27th Ann. Int’l Cryptology Conf. Advances in Cryptology (CRYPTO ’07), 2007.

[16] P. Golle, J. Staddon, and B. Waters, "Secure Conjunctive Keyword Search over Encrypted Data," Proc. Applied Cryptography and Network Security, pp. 31-45, 2004.

[17] L. Ballard, S. Kamara, and F. Monrose, "Achieving Efficient Conjunctive Keyword Searches over Encrypted Data," Proc. Seventh Int'l Conf. Information and Comm. Security (ICICS '05), 2005.

[18] D. Boneh and B. Waters, "Conjunctive, Subset, and Range Queries on Encrypted Data," Proc. Fourth Conf. Theory Cryptography (TCC), pp. 535-554, 2007.

[19] R. Brinkman, "Searching in Encrypted Data," PhD thesis, Univ. of Twente, 2007.

[20] Y. Hwang and P. Lee, "Public Key Encryption with Conjunctive Keyword Search and Its Extension to a Multi-User System," Pairing, vol. 4575, pp. 2-22, 2007.

[21] J. Katz, A. Sahai, and B. Waters, "Predicate Encryption Supporting Disjunctions, Polynomial Equations, and Inner Products," Proc. 27th Ann. Int'l Conf. Theory and Applications of Cryptographic Techniques (EUROCRYPT), 2008.

[22] A. Lewko, T. Okamoto, A. Sahai, K. Takashima, and B. Waters, "Fully Secure Functional Encryption: Attribute-Based Encryption and (Hierarchical) Inner Product Encryption," Proc. 29th Ann. Int'l Conf. Theory and Applications of Cryptographic Techniques (EUROCRYPT '10), 2010.

[23] E. Shen, E. Shi, and B. Waters, "Predicate Privacy in Encryption Systems," Proc. Sixth Theory of Cryptography Conf. Theory of Cryptography (TCC), 2009.

[24] M. Li, S. Yu, N. Cao, and W. Lou, "Authorized Private Keyword Search over Encrypted Data in Cloud Computing," Proc. 31st Int'l Conf. Distributed Computing Systems (ICDCS '10), pp. 383- 392, June 2011.

[25] C. Wang, N. Cao, J. Li, K. Ren, and W. Lou, "Secure Ranked Keyword Search over Encrypted Cloud Data," Proc. IEEE 30th Int'l Conf. Distributed Computing Systems (ICDCS '10), 2010.

[26] "A Complete Introspection on Big Data and Apache Spark" S. Praveen kumar, Dr. Y srinivas, Dr. D. SubaRao, Ashish Kumar.

[27] "A novel model for data leakage detection and prevention in distributed environment" S. Praveen kumar, Dr. Y srinivas, Dr. D. SubaRao, Ashish Kumar