Big Data in the Cloud

Sm Zobaed and Mohsen Amini Salehi

High Performance Cloud Computing (HPCC) laboratory,
University of Louisiana at Lafayette,
Lafayette, Louisiana 70503, USA
{*s.m.zobaed1 , amini*}*@louisiana.edu*

May 21, 2020

1 Overview

Cloud storage services have emerged to address the increasing demand to store and process huge amount of data, generally alluded as "Big Data". Typically, organizations store the huge volume of data to various clouds.

Cloud computing offers organizations the ability to manage big data and process them without the cost and burden of maintaining and upgrading local computing resources. However, efficient utilization of clouds for big data imposes new challenges in several domains. In this chapter, we discuss challenges in big data storage, distribution, security, and real-time processing. It is also explained how clouds can be instrumental for big data generated by Internet of Things (IoT). An overview of popular tools or services that are available in clouds for big data analytics is depicted. Finally, there is a discussion on future research directions in these areas.

2 key Research Findings Big Data in the Cloud

2.1 Storage Levels for Big Data in the Cloud

A cloud storage consists of enormous number (order of thousands) of storage server clusters connected by a high-bandwidth network. A storage middleware (*e.g.,* SafeSky (R. Zhao et al. 2015)) is used to provide distributed file system and to deal with storage allocation throughout the cloud storage.

Generally, cloud providers offer various storage levels with different pricing and latencies. These storage levels can be utilized to store big data in the cloud, depending on the price and latency constraints. Amazon cloud, for instance, provides *object storage, file storage,* and *block storage* to address the sheer size, velocity, and formats of big data.

Object storage is cheaper and slower to access, generally used to store large objects with low access rate (*e.g.,* backups and archive data). This type of storage service operates based on Hard Disc Drive (HDD) technology. Amazon simple storage service[1] (S3) is an example of Object storage service (J. Wu et al. 2010).

File storage service is also offered based on HDD storage technology. However, it provides file system interface, file system access consistency, and file locking. Unlike Object storage, file storage capacity is elastic which means growing and shrinking automatically in response to addition and removal of data. Moreover, it works as a concurrently-accessible storage for up to thousands of instances. Amazon Elastic File System[2] (EFS) is an example of such type of storage.

Block storage level is offered based on Solid-State Drives (SSD) that provide ultra-low access latency. Big data with frequent access or delay sensitive can be stored in this storage level. Amazon Elastic Block Store (EBS) is an example of Amazon cloud service offered based on block storage. High-performance big data applications commonly use these to minimize their access latency.

[1] https://aws.amazon.com/s3/
[2] https://aws.amazon.com/efs/

2.2 Storage-Efficient Cloud Big Data

Data could be replicated based on their demands of various forms. However, we would face storage and processing difficulty if we stored different possible forms of big data. Intelligent solutions can be applied to determine hot (popular) data points (*e.g.,* top viewed YouTube video) to replicate in several formats (Darwich et al. 2019). The remaining slight accessed data should be processed to replicate on-demand. This can be helpful to reduce big data storing cost and also ease the situation where we have storage limitation, particularly, edge nodes (Veillon et al. 2019).

2.3 Cloud versus Edge: Persistence versus Distribution

Although clouds provide an ideal environment for big data storage, the access delay (aka latency) to them is remarkable due to their centralized nature (Terzo et al. 2013). The delay can be particularly problematic for frequently accessed data (*e.g.,* index of big data search engine). As such, in addition to storage services, other cloud services need to be in place to manage distribution of big data so that the access delay is reduced. That is, separating the big data storage cloud services from the content distribution services.

Content Delivery Network (CDN) is a network of distributed servers aiming to reduce data access delay over the Internet (Pierre and Van Steen 2006). The CDN replicates and caches data within a network of servers dispersed at different geographical locations. Cloud providers offer distribution services through built-in CDN services or by integrating with current CDN providers. For instance, Amazon cloud offers CloudFront (Dignan 2008) CDN service. Akamai, a major CDN provider has been integrated with Microsoft Azure cloud to provide short delay in accessing big data (Sirivara 2016). Figure 1 depicts how CDN technology operates. As we can see in this figure, upon receiving a request to access data, CDN control server caches the data in the nearest edge server to deliver it with the minimum latency. Cloud providers offer distribution services through built-in CDN services or by integrating with current CDN providers. For instance, Amazon cloud offers CloudFront (Dignan 2008) CDN service. Akamai, a major CDN provider has been integrated with Microsoft Azure cloud to provide short delay in accessing big data (Sirivara 2016).

More recently, *fog/edge computing* paradigm has augmented the caching ability of CDN technology by integrating processing ability to it. To reduce latency and provide an uninterrupted service, Veillon *et al.*has developed a fog computing platform that processes the contents in a lazy (*i.e.,* on demand) manner for a particular geographical area. This is particularly useful for scenarios where multiple versions of the same content is required (*e.g.,* in video streaming platforms (Li et al. 2016)). However, contents that are highly demanded in that geographical area (*e.g.,* hot video contents (Darwich et al. 2019)) can be still cached in multiple formats by the fog server.

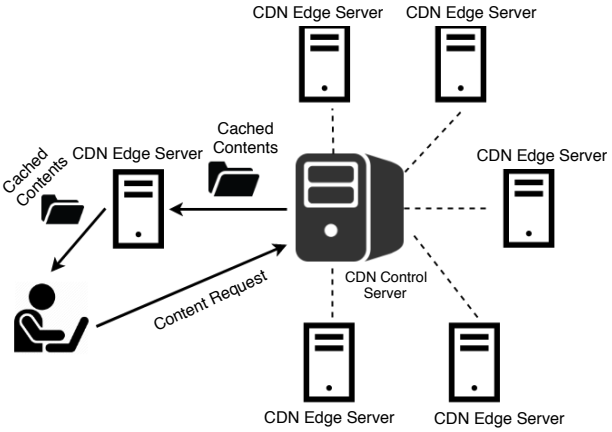

Figure 1: *CDN technology helps to minimize access latency for frequently-accessed big data in the cloud.*

2.4 Cloud and Big Data Security

One major challenge in utilizing cloud for big data is security concerns data owners. They do not have full control over their data hosted by clouds. The data can be exposed to external or, more importantly, to internal attacks.

A survey carried out by McKinsey (Arul Elumalai and Tandon 2016) showed that security and compliance are the two primary considerations in selecting a cloud provider, even beyond cost benefits. Recently, external attackers performed a Distributed Denial of Service (DDoS) attack that made several big data dealing organizations (*e.g.,* Twitter, Spotify, Github) inaccessible. Prism (Greenwald and MacAskill 2013) is an exmaple of an internal attack launched in 2013. In this section, we discuss major security challenges for big data in the cloud and current efforts to address them.

2.4.1 Security of Big Data Against Internal Attackers

In addition to common security measures (*e.g.,* logging or node maintenance against malware), user-side encryption(Salehi et al. 2014; Jason Woodworth et al. 2016) is becoming a popular approach to preserve the privacy of big data hosted in the cloud. That is, the data encryption is done using user's keys and in the user premises, therefore, external attackers cannot see or tamper data. However, user-side encryption brings about higher computational complexity and hinders searching big data(Zobaed et al. 2019).

Several research works have been undertaken recently to initiate different types of search over encrypted data in the cloud. Figure 2 provides a taxonomy of research works on the search over encrypted

data in the cloud. In general, proposed search methods are categorized as search over structured and unstructured data. Structured searches can be further categorized as SQL-aware, property-preserving, and format preserving encryption. Unstructured searches can be further categorized as keyword search, regular expression (regex) search, and semantic search.

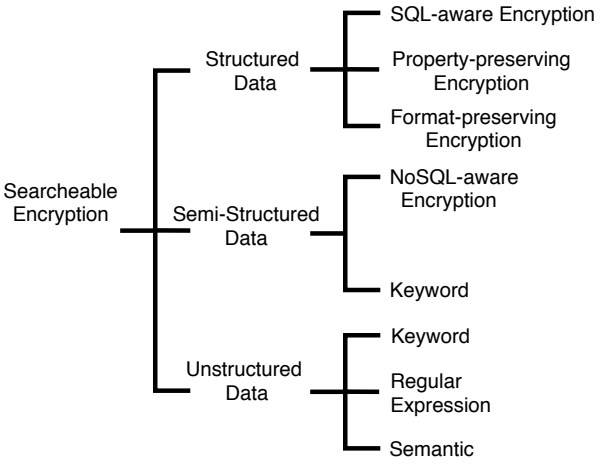

Figure 2: *Taxonomy of different types of search over encrypted big data in the cloud.*

Privacy-Preserving query over encrypted graph-structured data (Cao et al. 2011), cryptDB (Popa et al. 2012), and dragonfruit (Rozier et al. 2013) are the instances of search over encrypted structured data. SecureNoSQL (Ahmadian et al. 2017), SemiLD (Kettouch et al. 2019, and XSnippets (Naseriparsa et al. 2019) are the instances of search over encrypted semi-structured data. REseED (Salehi et al. 2014), SSE (Curtmola et al. 2006) S3C (Jason Woodworth et al. 2016) are tools developed for regular expression (Regex), keyword, and semantic searching respectively over unstructured big data in the cloud.

Big data integrity in the cloud is another security concern. The challenge is how to verify the data integrity in the cloud storage without downloading and then uploading the data back to cloud. Rapid incoming of new data to the cloud, makes verifying data integrity further complicated (D. Chen and H. Zhao 2012). Techniques such as continuous integrity monitoring are applied to deal with big data integrity challenge in the cloud (Lebdaoui et al. 2016).

2.4.2 Security Against External Attackers

External attackers are generally considered as potential threats to the cloud that handle big data. If a cloud face external attacks, user's data will be leaked and controlled by attackers. So, implementing security on cloud that handle big data is a challenge.

Researches have been done to address security issue. For instance, a layered framework (Reddy et al. 2012) is proposed for assuring big data analytics security in the cloud. It consists of securing multiple layers, namely virtual machine layer, storage layer, cloud data layer and secured virtual network monitor layer.

2.5 Real-time Big Data Analytics in the Cloud

Cloud big data can be subject to real-time processing in applications that detect patterns or perceive insights from big data. For instance, huge index structures need to be accessed to enable a real-time search service in S3C(Jason Woodworth et al. 2016). Other instances, for big data analytics are sensor, financial data , and streaming data (Hu et al. 2014).

The faster organizations can fetch insights from big data, the greater chance in generating revenue, reducing expenditures, and increasing productivity. The main advantage of separating storage from distribution is to help organizations to process big data in real-time. In most cases, such type of collected data are raw and noisy, thus, are not ready for analysis. Hence, a further processing is required before analysis.

2.6 Cloud-based Big Data Analytics Tools

Unlike conventional data that are either structured (*e.g.,* XML, XAML, and relational databases) or unstructured, big data can be a mixture of structured, semi-structured, or unstructured data (X. Wu et al. 2014). It is approximated that around 85% of total data comes from organization, are unstructured and moreover, almost all the individuals generated data are also unstructured (Pusala et al. 2016) . Such heterogeneous big data cannot be analyzed using traditional database management tools (*e.g.,* relational database management system (RDBMS)). Rather, a set of new cloud-based processing tools have been developed for processing big data (M. Chen et al. 2014).

NoSQL- NoSQL (generally referred to as "Not Only SQL") is a framework for databases that provides high-performance processing for big data. In fact, conventional relational database systems have several features (*e.g.,* transactions management and concurrency control) that make them unscalable for big data. In contrast, NoSQL databases are relieved from these extra features and lend themselves well to distributed processing in the clouds. These database systems can generally handle big data processing in real-time or near real-time. In this part, an overview of the NoSQL databases offered by clouds is explained.

- **Hadoop MapReduce–** MapReduce framework was published by google in 2005 (Dittrich and Quiané-Ruiz 2012). Hadoop is an open source implementation tool based on MapReduce framework developed by Apache (Vavilapalli et al.

2013). Hadoop is a batch data analytics technology that is designed for failure-prone computing environments. The specialty of Hadoop is that developers need to define only the `map` and `reduce` functions that operate on big data to achieve data analytics. Hadoop utilizes a special distributed file system, named Hadoop Distributed File System (HDFS), to handle data (*e.g.,* read and write operations). Hadoop achieves fault-tolerance through data replication. Amazon Elastic MapReduce (EMR)[3] is an example of Hadoop framework (Jourdren et al. 2012) offered as a service by Amazon cloud. Amazon EMR helps to create and handle elastic clusters of Amazon EC2 instances running Hadoop and other applications in the Hadoop system.

- **Key-Value Store–** key-value database system is designed for storing, retrieving, and handling data in mapping format where data are considered as a `value` and a `key` is used as an identifier of a particular data as because`keys` and `values` are resided pairwise in the database system. Amazon DynamoDB[4], GoogleBigTable[5], and HBase[6] are examples of cloud-based key-value store databases to handle big data.
- **Document Store–** These database systems are similar to key-value store databases. However, keys are mapped to documents, instead of values. Pairs of keys and documents (*e.g.,* in JASON or XML formats) are used to build data model over big data in the cloud. This type of database systems are used to store semi-structured data as documents, usually in JSON, XML or XAML formats. Amazon cloud offers Amazon SimpleDB[7] as a document store database service that creates and manages multiple distributed replicas of data automatically to enable high availability and data durability. Data owner can change data model on the fly, and data is automatically indexed later on.
- **Graph Databases–** These database systems are used for processing data that have graph nature. For instance, users' connections in a social network can be modeled using a graph database (Shimpi and Chaudhari 2012). In this case, users are considered as nodes and connections between users are represented as edges in the graph. The graph can be processed to predict links between nodes. That is, possible connections between users. Giraph [8], Neo4j[9], and FlockDB[10] are examples of cloud-based graph database services. Facebook, a large social networking site is currently using Giraph database system to store and analyze their users' connections.

In Memory Analytics- These are techniques that process process big data in the main memory (RAM) with reduced latency (Dittrich and Quiané-Ruiz 2012). For instance, Amazon provides ElasticCache[11] service to improve web applications' performance which at previous generally relied on slower disk-based databases. Beside that there are also tools for performing **in memory analytics** on the big data in cloud such as Apache Spark, SAP, and Microstrategy (Sultan 2015).

Different big data analytic tools introduced in this section are appropriate for different types of datasets and applications. There is no one perfect solution for all types of big data analytics in the cloud.

2.7 IoT Big Data in the Cloud and Edge

The Internet of Things (IoT) is a mesh of physical devices embedded with electronics, sensors, and network connectivity to collect and exchange data (Atzori et al. 2010). IoT devices are deployed in different systems, such as smart cities, smart homes, or to monitor environments (*e.g.,* smart grid). Basic work-flow of IoT systems that use cloud services (Dolgov 2017) is depicted in 3. According to the figure, the IoT devices produce streams of big data that are hosted and analyzed using cloud services. Many of these systems (*e.g.,* monitoring system)require low latency (real-time) analytics of the collected data. Big data generated in IoT systems generally suffer from redundant or noisy data which can affect big data analytic tools (M. Chen et al. 2014).

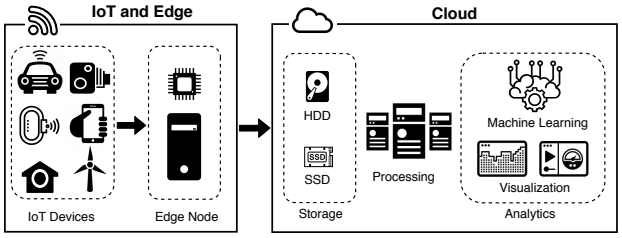

Figure 3: *Workflow of IoT systems, from big data generation and pre-processing them on the edge to big data storage and analytics on the cloud.*

Knowledge Discovery in Databases (KDD) and data mining technologies offer solutions for analysis of big data generated by IoT systems (Tsai et al. 2014). The big data are converted into useful information using KDD. The output of KDD are further processed using data mining or visualization techniques to extract insights or patterns of the data. Various tools such as AWS Greengrass, AWS IoT, Dell Statistica, Splunk, Pentaho, Azure stream analytics function based on the

[3]https://aws.amazon.com/emr/details/hadoop/
[4]https://aws.amazon.com/dynamodb/
[5]https://cloud.google.com/bigtable/
[6]https://hbase.apache.org/
[7]https://aws.amazon.com/simpledb/
[8]http://giraph.apache.org/
[9]https://zeroturnaround.com/rebellabs/examples-where-graph-databases-shine-neo4j-edition/
[10]https://github.com/twitter-archive/flockdb/
[11]https://aws.amazon.com/elasticache

aforementioned work-flow in clouds to extract insights from IoT big data with low latency (*Amazon IoT* 2015).

Fog and edge computing (Bonomi et al. 2012) extends the ability of clouds to edge servers with the goal of minimizing latency and maximizing the efficiency of core clouds. Fog computing can also be used to deal with noisy big data in IoT (ibid.). For instance, AWS enables creation of fog computing that can be pipelined to other Amazon cloud services (*e.g.,* Amazon SimpleDB and EMR) for big data processing.

2.8 Cloud-based Enterprise Search Services over Big Data

Providing access and search ability over big data is essential and data without these abilities is not much of use. However, organizations that deploy cloud services for their big data are concerned about data exposure (Woodworth and Salehi 2019; Zobaed et al. 2019). Hence, accessing the data without exposure is required.

Enterprise search cloud services are becoming increasingly popular to enable searching over and providing legitimate access to organizational big data (Kehoe 2009). Enterprise search services often maintain a dynamic index structure based on timely crawling in organizational documents. Then, the user's query is searched against the index structure and the result-set, referencing the relevant documents, is displayed to the legitimate user. Amazon cloud has provided a semantic enterprise search service named Kendra by leveraging machine learning and natural language processing methods. Amazon argues that their clients, such as Woodside, 3M, and Sage have improved the accuracy and speed of searching and accessing their organizational documents, in compared to other existing solutions(*Announcing Amazon Kendra: Reinventing Enterprise Search with Machine Learning* 2019).

We note that currently Amazon Kendra does not support enterprise search service over datasets encrypted by the user's key (aka user-side encryption). This leaves the organizational data privacy concern an open question in the cloud era. To address this concern, multiple solutions are provided to enable semantic search over user-side encrypted big data (Woodworth and Salehi 2019), (Zobaed et al. 2019), (Ahmad et al. 2019). These solutions aim at performing real-time search operation without compromising data privacy.

3 Future Directions for Research

3.1 Multi Source and Multi Cloud Big Data Processing Solutions

It can be said that the vast amount of data produced by business organizations or society have never been so large (X. Wu et al. 2014). Moreover, it is being generated from single or multiple sources and being stored in different clouds at a significant speed. Business industries want to get insights of big data quicker to dominate over other competitors by increasing profit per customer, reducing expenditures and optimizing operational steps. However, big data processing and analytics are still challenging jobs that require large computational system, costly software, and effort.

Based on the wide varieties of big data and analytics requirements, there is a large number of solutions available for data processing. Understanding business requirements are important to choose proper tools to access big data. Developments of APIs may explain the big data analytics problem briefly to the users and suggest solutions (Assunção et al. 2015).

Besides that, big data may come from multiple sources and there may be required real-time analysis. But if there is limited and fixed number of computational systems to process, processing will be impacted. So. cloud infrastructure should have more rapid elasticity capacity to handle the large amount of suddenly incoming data as quickly as possible.

3.2 Cloud Big Data and Serverless Computing Paradigm

A *serverless* computing (*e.g.,* AWS Lambda) allows developers to write and deploy code without managing or provisioning the deployment infrastructure (McGrath and Brenner 2017). Traditionally, a big data application developer had to deal with details of the required storage, database, and processing services for deployment. This entails knowledge and expertise in several fields, hence, the deployment process is slowed down. Alternatively, serverless paradigm removes the burden of scaling and other provisioning tasks from developers. Such paradigm relies on functions-as-a-service (FaaS), where developers divides their applications into small and stateless blocks to make them executable without any concern about the underlying cloud services (Carey 2019). Serverless paradigm also offers benefits from the incurred cost perspective. They charge their users only on their actual service usage time. That is, a user is not incurred any cost for the idle time of cloud resources.

4 Conclusion

The size of digital data is rapidly growing to the extent that storage and processing have become challenging. Cloud services have emerged to address these challenges. Nowadays, cloud providers offer collection of services that can address different big data demands. In particular, they provide levels of storage with different prices and access latency. They also provide big data distribution based on CDN techniques to reduce access latency further. For big data analytics, clouds offer several services, including NoSQL databases and in memory big data analytics. Although clouds have

been instrumental in enabling big data analytics, security and privacy of data are still major impediments for many businesses to embrace clouds. Additionally, Many small or medium organizations may not afford the cloud infrastructure cost. So, it can be said that dealing with big data is not well-suited for all types of organizations.

Bibliography

Ahmad, Sahan, SM Zobaed, Raju Gottumukkala, and Mohsen Amini Salehi (2019). "Edge computing for user-centric secure search on cloud-based encrypted big data". In: *Proceedings of 21st International Conference on High Performance Computing and Communications (HPCC)*. Guangzhou, China, pp. 662–669.

Ahmadian, Mohammad, Frank Plochan, Zak Roessler, and Dan C Marinescu (2017). "SecureNoSQL: An approach for secure search of encrypted NoSQL databases in the public cloud". In: *Jounal of International Journal of Information Management* 37.2, pp. 63–74.

Amazon IoT (2015). `https://aws.amazon.com/iot-platform/`. [Online; accessed 13-October-2017].

Announcing Amazon Kendra: Reinventing Enterprise Search with Machine Learning (2019). `https://aws.amazon.com/about-aws/whats-new/2019/12/announcing-amazon-kendra-reinventing-enterprise-search-with-machine-learning/`. [Online; accessed 16-February-2020].

Arul Elumalai, Irina Starikova and Sid Tandon (2016). *IT as a service: From build to consume*. `https://www.mckinsey.com/industries/high-tech/our-insights/it-as-a-service-from-build-to-consume/`. [Online; accessed 12-October-2017].

Assunção, Marcos D, Rodrigo N Calheiros, Silvia Bianchi, Marco AS Netto, and Rajkumar Buyya (2015). "Big Data computing and clouds: Trends and future directions". In: *Journal of Parallel and Distributed Computing* 79, pp. 3–15.

Atzori, Luigi, Antonio Iera, and Giacomo Morabito (Oct. 2010). "The Internet of Things: A Survey". In: *Journal of Computer Networks* 54.15, pp. 2787–2805.

Bonomi, Flavio, Rodolfo Milito, Jiang Zhu, and Sateesh Addepalli (2012). "Fog computing and its role in the internet of things". In: *Proceedings of the 1st edition of the MCC workshop on Mobile cloud computing,* MCC '12, pp. 13–16.

Cao, Ning, Zhenyu Yang, Cong Wang, Kui Ren, and Wenjing Lou (2011). "Privacy-Preserving Query over Encrypted Graph-Structured Data in Cloud Computing". In: *Proceedings of the 31st International Conference on Distributed Computing Systems*. ICDCS '11. Washington, DC, USA, pp. 393–402. ISBN: 978-0-7695-4364-2.

Carey, Scott (2019). *What is serverless computing and which enterprises are adopting it?* `https://www.computerworld.com/article/3427298/what-is-serverless-computing-and-which-enterprises-are-adopting-it.html`. [Online; accessed 1-March-2020].

Chen, Deyan and Hong Zhao (2012). "Data security and privacy protection issues in cloud computing". In: *Proceedings of International Conference on Computer Science and Electronics Engineering*. Vol. 1. ICCSEE '12, pp. 647–651.

Chen, Min, Shiwen Mao, and Yunhao Liu (2014). "Big data: A survey". In: *Journal of Mobile Networks and Applications* 19.2, pp. 171–209.

Curtmola, Reza, Juan Garay, Seny Kamara, and Rafail Ostrovsky (2006). "Searchable symmetric encryption: improved definitions and efficient constructions". In: *Proceedings of the 13th ACM conference on Computer and communications security*. CCS '06, pp. 79–88.

Darwich, Mahmoud, Mohsen Amini Salehi, Ege Beyazit, and Magdy Bayoumi (2019). "Cost-efficient cloud-based video streaming through measuring hotness". In: *Journal of The Computer* 62.5, pp. 641–656.

Dignan, Larry (2008). *Amazon launches CloudFront; Content delivery network margins go kaboom*. `http://www.zdnet.com/article/amazon-launches-cloudfront-content-delivery-network-margins-go-kaboom/`. [Online; accessed 13-October-2017].

Dittrich, Jens and Jorge-Arnulfo Quiané-Ruiz (2012). "Efficient big data processing in Hadoop MapReduce". In: *Journal of the VLDB Endowment* 5.12, pp. 2014–2015.

Dolgov, Sergey (2017). *AI Marketplace: Neural Network in Your Shopping Cart*. `https://www.linkedin.com/pulse/ai-marketplace-neural-network-your-shopping-cart-sergey-dolgov-1/`. [Online; accessed 13-October-2017].

Greenwald, Glenn and Ewen MacAskill (2013). "NSA Prism program taps in to user data of Apple, Google and others". In: *Journal of The Guardian* 7.6, pp. 1–43.

Hu, Han, Yonggang Wen, Tat-Seng Chua, and Xuelong Li (2014). "Toward scalable systems for big data analytics: A technology tutorial". In: *Journal of IEEE access* 2, pp. 652–687.

Jourdren, Laurent, Maria Bernard, Marie-Agnès Dillies, and Stéphane Le Crom (June 2012). "Eoulsan". In: *Journal of Bioinformatics* 28.11, pp. 1542–1543.

Kehoe, Miles (2009). *Mapping Security Requirements to Enterprise Search - Part 1: Defining Specific Security Requirements*. `http://www.ideaeng.com/security-eprise-search-p1-0304`. [Online; accessed 16-February-2020].

Kettouch, Mohamed, Cristina Luca, and Mike Hobbs (2019). "SemiLD: mediator-based framework for keyword search over semi-structured and linked data". In: *Journal of Intelligent Information Systems* 52.2, pp. 311–335.

Lebdaoui, Imane, Said El Hajji, and Ghizlane Orhanou (2016). "Managing big data integrity". In: *Proceed-*

ings of International Conference on Engineering & MIS. ICEMIS '16, pp. 1–6.

Li, Xiangbo, Mohsen Amini Salehi, Magdy Bayoumi, and Rajkumar Buyya (2016). "CVSS: A Cost-Efficient and QoS-Aware Video Streaming Using Cloud Services". In: *Proceedings of the 16th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*. CCGrid '16. IEEE, pp. 106–115.

McGrath, Garrett and Paul R Brenner (2017). "Serverless computing: Design, implementation, and performance". In: *Proceedings of 37th International Conference on Distributed Computing Systems Workshops (ICDCSW)*, pp. 405–410.

Naseriparsa, Mehdi, Md Saiful Islam, Chengfei Liu, and Lu Chen (2019). "XSnippets: Exploring semi-structured data via snippets". In: *Journal of Data & Knowledge Engineering* 124, p. 101758.

Pierre, Guillaume and Maarten Van Steen (2006). "Globule: a collaborative content delivery network". In: *Journal of communications Magazine* 44.8, pp. 127–133.

Popa, Raluca Ada, Catherine MS Redfield, Nickolai Zeldovich, and Hari Balakrishnan (2012). "CryptDB: processing queries on an encrypted database". In: *Journal of Communications of the ACM* 55.9, pp. 103–111.

Pusala, Murali K, Mohsen Amini Salehi, Jayasimha R Katukuri, Ying Xie, and Vijay Raghavan (2016). "Massive Data Analysis: Tasks, Tools, Applications, and Challenges". In: *Big Data Analytics*. Springer, pp. 11–40.

Reddy, C Kishor Kumar, PR Anisha, K Srinivasulu Reddy, and S Surender Reddy (2012). "Third party data protection applied to cloud and XACML implementation in the hadoop environment with sparql". In: *Journal Of International Organization of Scientific Research of Computer Engineering (IOSRJCE)* 2.1.

Rozier, Eric WD, Saman Zonouz, and David Redberg (2013). "Dragonfruit: Cloud provider-agnostic trustworthy cloud data storage and remote processing". In: *Proceedings of 19th Pacific Rim International Symposium on Dependable Computing*. IEEE, pp. 172–177.

Salehi, Mohsen Amini, Thomas Caldwell, Alejandro Fernandez, Emmanuel Mickiewicz, Eric WD Rozier, Saman Zonouz, and David Redberg (2014). "RESeED: regular expression search over encrypted data in the cloud". In: *Proceedings of 7th International Conference on Cloud Computing*, pp. 673–680.

Shimpi, Darshana and Sangita Chaudhari (2012). "An overview of graph databases". In: *Proceedings of the 2nd International Conference in Recent Trends in Information Technology and Computer Science*. ICRTITCS '12, pp. 16–22.

Sirivara, Sudheer (2016). *Windows Azure Content Delivery Network*. `https://azure.microsoft.com/en-us/blog/azure-cdn-from-akamai-ga/`. [Online; accessed 13-October-2017].

Sultan (2015). *Top 10 In-Memory Business Intelligence Analytics Tools*. `https://www.mytechlogy.com/IT-blogs/9507/top-10-in-memory-business-intelligence-analytics-tools/`. [Online; accessed 12-October-2017].

Terzo, Olivier, Pietro Ruiu, Enrico Bucci, and Fatos Xhafa (2013). "Data as a service (DaaS) for sharing and processing of large data collections in the cloud". In: *Proceedings of the 7th International Conference on Complex, Intelligent, and Software Intensive Systems*. CISIS '2013, pp. 475–480.

Tsai, Chun-Wei, Chin-Feng Lai, Ming-Chao Chiang, Laurence T Yang, et al. (2014). "Data mining for Internet of Things: A survey." In: *Journal of IEEE Communications Surveys and Tutorials* 16.1, pp. 77–97.

Vavilapalli, Vinod Kumar et al. (2013). "Apache Hadoop YARN: Yet Another Resource Negotiator". In: *Proceedings of the 4th Annual Symposium on Cloud Computing*. SOCC '13. New York, NY, USA, 5:1–5:16. ISBN: 978-1-4503-2428-1.

Veillon, Vaughan, Chavit Denninnart, and Mohsen Amini Salehi (2019). "F-FDN: Federation of Fog Computing Systems for Low Latency Video Streaming". In: *Proceddings of the 3rd International Conference on Fog and Edge Computing (ICFEC)*, pp. 1–9.

Woodworth, J and Mohsen Amini Salehi (2019). "S3BD: Secure semantic search over encrypted big data in the cloud". In: *Concurrency and Computation: Practice and Experience* 31.11.

Woodworth, Jason, Mohsen Amini Salehi, and Vijay Raghavan (2016). "S3C: An architecture for space-efficient semantic search over encrypted data in the cloud". In: *Proceedings of International Conference on Big Data*. Big Data '16, pp. 3722–3731.

Wu, Jiyi, Lingdi Ping, Xiaoping Ge, Ya Wang, and Jianqing Fu (2010). "Cloud storage as the infrastructure of cloud computing". In: *Proceeding of 9th International Conference on Intelligent Computing and Cognitive Informatics*. ICICCI '10, pp. 380–383.

Wu, Xindong, Xingquan Zhu, Gong-Qing Wu, and Wei Ding (Jan. 2014). "Data Mining with Big Data". In: *Journal of IEEE Transactions on Knowledge and Data Engineering* 26.1, pp. 97–107.

Zhao, Rui, Chuan Yue, Byungchul Tak, and Chunqiang Tang (2015). "SafeSky: A Secure Cloud Storage Middleware for End-User Applications". In: *Proceedings of the 34th IEEE Symposium on Reliable Distributed Systems*. SRDS '15, pp. 21–30.

Zobaed, SM, Sahan Ahmad, Raju Gottumukkala, and Mohsen Amini Salehi (2019). "Clustcrypt: Privacy-preserving clustering of unstructured big data in the cloud". In: *Proceedings of 21st International Conference on High Performance Computing and Communications; (HPCC)*. Guangzhou, China, pp. 609–616.