# Big Data Key Challenges

**Sultan Alotaibi**

College of Computer and Information Systems, Umm Al-Qura University, Makkah, 24381 Saudi Arabia

**Summary**
The big data term refers to the great volume of data and complicated data structure with difficulties in collecting, storing, processing, and analyzing these data. Big data analytics refers to the operation of disclosing hidden patterns through big data. This information and data set cloud to be useful and provide advanced services. However, analyzing and processing this information could cause revealing and disclosing some sensitive and personal information when the information is contained in applications that are correlated to users such as location-based services, but concerns are diminished if the applications are correlated to general information such as scientific results. In this work, a survey has been done over security and privacy challenges and approaches in big data. The challenges included here are in each of the following areas: privacy, access control, encryption, and authentication in big data. Likewise, the approaches presented here are privacy-preserving approaches in big data, access control approaches in big data, encryption approaches in big data, and authentication approaches in big data.

***Key words:***
*Access Control, Big data, Privacy, Authentication.*

## 1. Introduction

Big data has become a centralized and significant topic in contemporary business and science. Big data refers to massive data sets that are more complex and complicated structures and the operations of processing, analyzing, or storing are challenges. These data can be videos, pictures, medical records, social networking activities, queries, search outcomes, sensors data, mobile phone applications data, geographic information systems data, scientific results data, audios, e-commerce transactions, emails, posts, or logs ..etc. Those data are stored in databases. The databases containing these data sets are incredibly growing and increased, so controlling and managing these data sets has become very difficult.

Till 2003, the amount of data is created is around 5 Exabytes. Nowadays, however, this amount of data needs only two days to be generated. In the future, the amount of data is expected to be tens of Zettabytes. For example, one million servers are distributed around the world for Google. Moreover, the number of mobile subscriptions is reach six billion, and around ten billion messages are exchanged per day. Also, the number of devices connected to the Internet might reach 100 billion by 2025. 140 billion images are uploaded on Facebook and 48 hours of videos are uploaded. Some properties are associated with big data and big data is characterized by: Variety, Volume, and Velocity.

The variety of the data being produced is categorized into three types structured, semi-structured, and unstructured data. Unstructured data is hard to manage and analyze while structured data is simply stored and recovered. The semi-structured data consists of tags that can detach elements of the stored data. The volume of big data exceeds zettabyte. This size of data has become a serious challenge for being handled using traditional systems. Velocity in big data refers to the speed of coming data from diverse sources as well as which data is in the flow. The traditional systems are unqualified to execute data analytics in data under motion conditions such as sensors data [1] [2].

The main concerns of big data are security and privacy issues. Privacy becomes a serious issue when the applications of big data contain sensitive and personal information are considered. However, some big data applications do not contain sensitive or personal information such as astronomy or e-science application. Applications based on social web or business analytics are privacy critical domains; for example. The privacy concern in big data content revolves around what and how the companies governing processing this data are doing with this data.

### 1.1 Privacy Requirements

The main concern with big data applications is that they are not using data for only the purpose of their storage. Big data analytics purposes are for processing data stored in big data to reveal hidden patterns and correlations of the information. With improved big data analytics invasion of privacy has become easy and potential. Moreover, unknown provenance of data has become another challenge with no implementation of authenticity and integrity on processed data.

Manuscript received April 5, 2022
Manuscript revised April 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.4.40

- Requirements in big data collection: the possibility of eavesdropping on the collected data is aroused when the data is being sensitive and personal data. Therefore, mechanisms of protecting the data are needed as well as security techniques for ensuring the privacy of the personal data.
- Requirements in big data storage: more than particular individual personal data would be targeted and disclosed when the confidentiality of information is not ensured and considered.
- Requirements in big data processing: processing the data is considered a major key component for big data analytics. Therefore, privacy becomes a serious concern when big data processing by analytics.

### 1.2 Big data – Big benefits

Big data has attracted big industries. Effective data management strategy leads to effective financial performance. Enterprises, manage the data effectively, stand out longer and grow faster. McKinsey Global Institute (MGI) has shown the effect of transforming entire sectors varying from healthcare to manufacturing to retails toward big data [3]. Big data assists industries increase productivity, so the demand for big data has been increased. In this section examples of big data, benefits are presented. The big data benefits should be minded when the potential risks that might impact an individual's privacy are considered. However, the risks associated with privacy should not discredit the benefits of big data. Also, with the current market conditions, the benefits of big data do not necessarily be accounted for by the individuals whose information is being collected.

(i) Healthcare
In [3], the researchers discovered terrible side effects when the patients take Pravachol, a drug used for reducing cholesterol, and Paxil, a popular antidepressant prescription. The side effect of using both medicines together is to increase the glucose which is included in the blood and its degree can reach the diabetic level. The researchers have completed their study using the data mining techniques that are used to determine the patterns in large datasets. They conducted their statistical analysis on the data stored in the Adverse Event Reporting System (AERS). They used a novel signal detection algorithm.

(ii) Smart grid
A smart grid is a new way to introduce the existing electricity grid by the concept of bi-directional movement of data and electricity. The smart gird controls the electricity usage by monitoring the usage itself. Big data helps the smart grid to provide the best power quality and effective distribution of the electricity with transferring toward suitable implementation of renewable energy.

(iii) Traffic management
Big data solutions can be used for traffic management and control. Toll-way pricing systems have been considered by governments around the world. The charges differ based on congestion and mobility. The personal location information facilities the decisions of road construction or diminishing of the traffic for the city planners and developing urban decision-makers. Likewise, motorists can benefit from the smart road information based to avoid congestion and traffic. Also, the navigation systems associated with vehicles are linked with communication components that provide services based on intelligent techniques to facilitate the trip.

(iv) Retail
The retail markets, also, are affected by big data. For example, the Retail Link system was invented and provided by Wall-Mart. The invented management system allows the providers to monitor the quantity of the products for each store at any time. Also, the feature "Customer Who Bought This Also Bought" which is shown after purchasing from the Amazon website motivates purchasers to buy further purchases which are suggested by intelligent filtering tools.

(v) Payments
Another effective benefit is when big data is used in detecting the deception in the card payment gateway. An efficient technique is needed for detecting uncommon transactions completed the first time by using the card. "Card-Not-Present transactions" is a solution that provides predictive fraud possibilities. This is performed by analyzing the buyers' purchases histories in real-time and according to the evaluation, the transaction will be considered as a fraudulent transaction or not.

### 1.3 Challenges in Big Data Analysis

Heterogeneity and incompleteness of the data are challenges in big data. The computer analysis algorithms and mechanisms always look forward to harmonious data. An appropriate structure of the data is an essential requirement by traditional data analysis techniques. In general, storing and processing information that is all congruent in volume and structure is most preferable by computer systems. Semi-structured data is more mandatory effort and work. After analyzing the data, some incompleteness data and errors remain. Therefore, this must be managed and controlled during the process of analyzing the data.

The big data term indicates a massive amount of data which is forming a challenge when managing and processing this size of data. The large data size was addressed by speeding up the processors. However, the data size has stretched faster and rapidly and the processors' speed remains static.

The greater amount of data sets that are processed arouses time consumption concern due to huge data requiring a longer time to be analyzed. Designing efficient systems should not only consider the ability to handle large data sets but also should consider analyzing the large data sets faster. For example, with a fraudulent credit card transaction, fast actions must be conducted to cancel the transaction before it is completed.

The privacy of the data is a significant challenge. For example, location-based services mandate oblige the subscribers to share their locations with the service providers and this arouses recognizable privacy concerns even if with hiding the users' identities.

## 2. Security and Privacy Challenges in Big Data

In this section, the security and privacy challenges in big data are introduced. The challenges in privacy-preserving, access model challenges, encryption challenges, and authentication challenges.

### 2.1 Privacy Preserving challenges in Big Data

The amount of data produced by social networks, medical systems, or even surveillance systems has been increased incredibly, so managing and storing this amount of data is not an easy task. Cloud computing is considered one of the smart solutions for big data. However, the privacy and security concerns challenge are rising and make it a less attractive alternative. A huge amount of images used in social networks or medical record systems might involve sensitive and personal information that can be revealed and disclosed. Moreover, insider attacks might occur by cloud services provider (CSP) itself because the provider has full control functionality of saved data. Encrypted data was considered a solution to avoid invasion the privacy attack. However, overhead computations need to be considered especially with encrypting the images, so traditional cryptography mechanisms and techniques are inefficient for preserving privacy in big data.

Processing large scale data is a serious challenge with the rapidly increasing speed and throughput of applications [4]. The problem would be more complicated when it comes to protected and encrypted data. Also, processing large scale data might reveal sensitive and personal information and makes them easy targets for attackers. The variety and massive volume of data complicate designing an appropriate and efficient protocol for preserving privacy. In addition, keeping up with rapid increases in data volume is being an obstacle for existing mechanisms dealing with privacy preservation in big data.

A similarity detection algorithm is used to analyze the data of users and detect the similarities between objects of a particular user [4]. This invasion of privacy is used for various purposes. For example, it is used to make a better decision for behavioral advertisements while the interests of users are revealed. Many applications improve their quality and performance at the expense of disclosing the users' sensitive information. The mechanisms of clustering when similar objects are grouped in the same cluster facilitate the operation of the similarity detection algorithm.

The demand for healthcare services has been increased incredibly and makes healthcare services expensive. Healthcare industries tend for digitizing medical records to improve processing and managing the data. Digitizing medical records increases the complexity of managing and processing these massive data. Big data is considered a promised solution to facilitate the processing and managing diversity of massive medical data. This transformation, also, can lower the cost of healthcare and improve the growth of the economies. However, big data in healthcare aroused patient privacy concerns. Centers that store healthcare data are tending to get HIPPA certification to ensure patients' records' safety. HIPPA confirms the security policies rather than implementing them. Also, the absence of laws and uncontrolled use of big data analytics causes the invasion of patient privacy [5].

Collected data is targeted by companies such as Facebook, Flicker, Google…etc. The main concern with this is the unawareness of the user of processing this data and applying different big data analytics to extract sensitive and personal information related to the users. Moreover, the potential threats that can occur from other users' uploading are a serious concern for users' privacy [6].

Big data analytics in collections of sensors data can reveal private data [7]. One obvious example of sensors data collection is sensors placed at smart homes. Sharing and storing personally identifiable data brings privacy concerns. The issues related to privacy invasion of personally identifiable data can be categorized into four main categories of concerns. The sensitivity of the data generated by sensors placed at smart homes is considered a concern because the ways of collecting these data are unknown. Also, using the insecure channel for transmitting these data violate the confidentiality and integrity of the information. Using untrusted and external storage delivers a threat to private information. Accessing the data without enforcing a proper Authentication and authorization process causes disclosing identifiable information.

### 2.2 Access Control Model Challenges in Big Data

Accessing the available resources should be limited according to the principle of least privilege. The various data types and huge data volume are an appropriate environment for occurring cybercrimes. The risks can be insider or outsider. Unfortunately, this might cause disclosing of sensitive and personal information. Therefore, limiting the number of persons who can access the data should be considered to provide secured data and ensure the integrity of the data. Also, intensive monitoring of these accesses is a significant procedure to ensure the security of the data.

According to [1], encrypting data based on access control policies need to be considered to guarantee to provide secure private and sensitive data. The approach of attribute-based encryption (ABE) needs to be improved and enhanced because it is lack efficiency and scalability. Also, validating the authenticity and fairness among spread entities are subjects that should be considered inefficient cryptography secure mechanisms based on access control policies. The public key cryptosystem is the basic of attributes-based encryption (ABE). However, some personal and sensitive information considered as least sensitive is not encrypted and this is still instrumental for analytics. Therefore, a comprehensive and efficient cryptographically secure framework needs to be implemented.

Systems and tools for information dissemination enable users to post their preferred lists of documents and topics. To subscribe to the systems, the users need to provide their interested topics. When new data are ready to be disseminated, systems will inform users who might be careful about them according to their interests. The main drawback of the current data dissemination systems is the absence of defined access control mechanisms. Controlling the access to data relies on the administrator because the administrator draws and determines the rules of accessing data [8]. For example, the administrator determines which users can access which data. Also, authorization and authentication are not considered which makes it unreliable.

### 2.3 Encryption Challenges in Big Data

In [9], data and information of big data can be efficiently handled by MapReduce. MapReduce is a big data solution example dealing with new arrives data in various forms. MapReduce is generated by Google and its concept is about a programming framework based on distributed computing idea and divide and conquer technique used to break down big data problems into subproblems as an attempt to lower the level of big data complexity [1]. However, its intermediate data is unprotected very well. Also, it does not support operations on ciphertexts. Therefore, the power of using MapReduce must not be stopped at those challenges. Data intermediate needs to be protected. Also, some operations and computations need to be supported over encrypted intermediate data. Conventional data encryption techniques can be used to protect the data. However, they complicate computations and operations over the protected and decrypted data.

In [10], encryption schemes were presented and discussed. The rapid improvement of the cloud paradigm stimulates the entire IT industry to adopt cloud based services and infrastructure due to its advantages such as flexibility and dropping cost. However, the integrity of the data as well as the confidentiality of the data has become major concerns. The stored data must be kept classified, whether the data belongs to the clients or organizations. Also, the data must not be modified or disclosed in any unauthorized way. Therefore, an efficient encryption method is needed to address this issue.

Based on [11], while cloud computing and big data have dramatically been improved, these enhancements change the shape of contemporary industries of information technology. The security of cloud computing and big data issues still need to be addressed and resolved. Encrypting data solutions that used a secure key to encrypt and decrypt data is not an efficient approach because it provides heavy computations on the client-side. Also, it is not possible to let the cloud provider encrypt and decrypt the data for security reasons. The issue will become more complex when this data is sharable. The proxy re-encryption (PRE) mechanism has been proposed in the literature. The concept of the PRE mechanism is to give the right of decrypting the data to the other users who are sharing the data with the originator user. The data cannot be decrypted by the cloud while the other user could decrypt the data by using their private key. This arouses a security concern regarding privacy protection. The problem is known as an abuse of re-encryption concern. The possibility of invasion the privacy will occur when a new member has delegated the re-encrypt key then the new member could decrypt old information that might not be authorized to access.

### 2.4 Authentication Challenges in Big Data

According to [12], multi-factor authentication (MFA) is a technique used to validate the user's login to the system. The MFA usually combines two or more authentication factors to validate the user. The MFA systems have been used commonly in the cloud system. Also, numerous academic investigations were proposed various MFA system techniques. The user, usually, is validated by password as the first authentication factor and the second or third authentication factor might be fingerprint or smartcard; for example. However, sensitive information might be exposed to untrusted cloud servers. Physical identification factors of the users might be kept and then used by the untrusted cloud servers later.

## 3. Security and Privacy Approaches in Big Data

In this section, the security and privacy approaches in big data are introduced. The approaches cover privacy preserving, access model, encryption, and authentication.

### 3.1 Privacy Preserving Approaches in Big Data

In this subsection, a framework is presented for maintaining privacy in collecting information from data sensors of smart homes. By 2050, the number of senior people in industrialized nations is going to be double the current number. Therefore, the need for professional people who provides healthcare services must be increased at least double the current number. Due to the environment of private homes being more undisturbed and convenient, elderly people prefer to stay at home. To meet this willingness, Aging-in-Place (AIP) uses sensor networks to provide healthcare services. Gathering data must be done centrally to deliver assistive services easily. The sensitivity of the collected data is a serious challenge. The data gathered from the sensors placed at smart homes represents personal data. Encrypted data is considered as a solution to keep sensitive data confidential. However, the encryption alternative is expensive and adds complexity.

In [7], an approach was presented to address the security and privacy concerns related to collected sensor data from smart homes. The proposed framework maintains the privacy of the data in three different lifecycles of the data (collection, storage, and processing). Also, the proposed technique classifies the data into three different areas (ownership of the data, transferring the data, storing the data, and accessing the data).

Also, in [10], the architecture of the proposed solution is presented and discussed. The architecture contains three main modules. The components of the architecture are the data collector, data receiver, and result provider.

The data collector is placed at each smart home. The main task of the data collector is to collect the data from the sensors and pass the data to the data cluster. The collected data must be confident when it is passed to the clusters. The SSL protocol is used to ensure the confidentiality of the transferred data. Also, SSH is used and considered to ensure the security of the transferred data and provide high-speed transfer. The second module is the data receiver. The data receiver module is receiving the transferred data from the data collector and takes one of two actions based on the algorithmic transformation function which categorizes the attributes into three categories (regulations, empirical observations, and linkage to public sources). The goal of implementing this algorithmic function is to separate the data into sensitive data and de-sensitized data. Therefore, the data receiver consists of two datasets. The real data with primary/quasi– identifiers values are hashed and stored in a de-identified sensor data unit. Once the hashed and real values from the set of primary/quasi- identifiers do not exist, the identifier data dictionary unit is used to store them.

Until here, the concerns regarding the secure collecting, storing, and processing of sensitive data are addressed. However, accessing the data need to be maintained and controlled. Therefore, the third module included in the proposed architecture is the result provider module. This module ensures preserving privacy and authenticating end users who might utilize data to improve healthcare services. Moreover, the access control module attached to this architecture must validate the authorizations for accessing the data. There are four categories (access control, identifier retrieval, transformer, and result processor) according to the potential activities. The level of privacy and the authentications as well as the authorizations are determined in the access control module. A list of personal/quasi-identifiers are generated based on the authorized access by the identifier retrieval module. This list is generalized/suppressed the real personal/quasi- identifiers in the transformer module and a dataset is created for hashed, real, and generalized/suppressed values. Based on the outputs of the transformer module, the result processor module acts. The last module which is the result processor changes the hashed personal/quasi- identifiers values by generalized/suppressed values according to the transformer module's results.

In this proposed scheme, the k- anonymization algorithm is used based on the level of the privacy of the end-users who aim to access the data for replacing values of de-identified store units to preserve the privacy of any results retrieved.

End-users and businesses use cloud computing to ease the computational resources and benefit from a low cost advantage. Cloud services can be Infrastructure as Service (IaaS), Platform as Service (PaaS), or Software as Service (SaaS). However, security issues related to cloud computing make the lack of interest shift to cloud computing. The ability to analyze the users' data makes the cloud vulnerable because that violates their privacy. There are many techniques and mechanisms used to analyze the users' data and to obtain valued and sensitive information. For example, the ads and recommended search results in google engine search are extracted according to the users' behavior. These techniques might be used by attackers to reveal sensitive and personal information. Storing all data of a single user in one cloud provider causes serious threats when data mining algorithms are used. Therefore, the provider can reveal sensitive and personal information by using advanced techniques and algorithms. All that the mining algorithms need is a practical aggregation of data to extract desired information. The purposes of attackers meet with the weakness and vulnerability of single cloud storage provider architecture.

Data mining techniques are used to disclose and discover hidden patterns from massive databases. The techniques and algorithms of data mining require huge data sets. Data mining techniques are used by cloud computing providers to offer improved services for clients. The clients might be unaware while their sensitive and personal information is analyzed by cloud computing providers, and this is a serious privacy invasion event. Moreover, the attacker might be authorized to access the cloud and can use data mining techniques to reveal others' sensitive and personal information. The successful data mining event for extracting useful information depends on the following elements first is an adequate amount of data and the second is appropriate mining algorithms and techniques.

In [13], an approach was proposed to protect privacy from data mining-based attacks. The architecture of the system and its components are given in detail in [13].

The system architecture contains two main components that are cloud data distributer and cloud providers. After the cloud data distributor receives data from the clients, it is splitting the data into chunks and allocates these chunks to different cloud providers. Then, the cloud providers save those chunks and reply when the chunks are retrieved.

(i) Cloud data distributor

The cloud data distributer entity receives the client data formed and segments the file into chunks and distributes these chunks across different cloud providers. Also, this entity is used when desired data is retrieved. It is, also, receives requests from the client and forwards these requests to the cloud providers. The cloud data distributer is placed between the client and the cloud providers, so the client does not network directly with cloud providers. The cloud data distributer is considered as an agent that represents the clients. When the client uploads the data/files into the cloud data distributer, each file is going to be given a privacy level. The client defines the privacy level for each file according to the mining sensitivity of each file. The mining sensitivity means the importance of the information that can be revealed when data mining algorithms run over the file. There are four different privacy levels (PL) are considered in this system: 0, 1, 2, and 3. Those levels are described and categorized according to the following categories: public data, low sensitive data, moderately sensitive data, and highly sensitive data/private data. As long the data is very sensitive as privacy level is going to be increased. The cloud data distributer entity receives files with their privacy levels and segments the files into chunks with the same privacy level and informs the client of the total number of the chunks. The number of chunks is needed for retrieving purposes. To conceal the identity of the client from the cloud provider each chunk is given a unique virtual id. Also, each cloud provider has been recognized by cloud data distributors by one of four privacy levels same as the file's privacy levels. The privacy levels of the cloud providers are assigned according to the reliability of each cloud provider. Therefore, the most reliable cloud provider is the most trusted cloud provider, and it is going to be assigned a higher privacy level. Moreover, the cost of the cloud provider is considered. For example, if two cloud providers have the same privacy level with different costs. The cloud data distributer is going to assign chunks to the cloud provider at a low cost.

The distribution of data and retrieving of data operations are performed by the cloud data distributor. To do those tasks cloud data distributer maintains associated data to clients, chunks, and cloud providers. Cloud data distributer controls three types of tables which are client table, chunk table, and cloud provider table. The client table consists of data associated with the client. The chunk table contains information associated with the chunks. The cloud provider table consists of entries regarding particular cloud provider information.

(ii) Cloud provider

The second entity of the system architecture is the cloud providers where the data is stored. The cloud provider is responsible for saving the received chunks, acting when chunks are requested, and deleting chunks when a removal request is made. The virtual id is the key to doing those tasks.

In terms of implementing the proposed system architecture, three functionalities are needed to be considered.

1) Distribute data
2) Retrieve data
3) Remove data

Two main functions are included in the cloud data distribution to maintain the distribution of the data.

- *Chunks [ ] split (file*) function receives the files from the users and splits the received file into fixed sixes chunks. Also, the virtual id is given to the chunk in this step.
- *Void distribute (chunks [ ]) is* considered, after splitting data into chunks and distributing the chunks for appropriate cloud provider storage.

Also, the cloud data distributor entity includes the functions responsible for retrieving the data and the functions are given as the following:

- *Chunk get_chunk (client name, password, filename, so no.)* function provides the ability to request chunks by the client and provides them to the client after obtaining them from the cloud provider where they are stored.
- *Chunks [ ] get_file(client name, password, filename)* function provides the ability to request the file by the client and provides the files to the client after it fetches the corresponding chunks associated with the requested file.

The following functions are used when removal of the data is desired:

- *Remove_chunk(client name, password, filename, so no.)* function receives chunk removal requests from the client and passes them to the cloud provider.
- *Remove_file(client name, password, filename)* function is responsible for file removal request.

The Internet of Things (IoT) combines various sensors and various domains of services and applications. A diverse of applications are developed to meet personalized needs and clients' requirements. Finding a suitable platform in an IoT environment to ease the interoperation through the diversity of the applications domains has become challenging. Also, The development of the Internet of Things (IoT) faces several obstacles such as optimizing energy consumption in sensors and controlling the massive and huge volume of data with providing security and privacy. Preserving privacy in diverse and shared data sources from sensed observations has become an issue in the IoT environment. In this subsection, a proposed mechanism from [14] is going to be represented. The proposed scheme is a negotiation-based privacy preserving scheme.

The utility of microdata causes an obstacle when privacy preservation is implemented in IoT. The microdata facilitates providing various useful services. Therefore, a certain amount of data is need to be considered to handle diverse kinds of computations. Therefore, most applications deal with utility as the main requirement while privacy comes as an accompanying restricting factor. The provided services will be affected when the privacy is integrated by protecting the microdata. The negotiation-based mechanism approach is considered to overcome this issue. Both the utility and the privacy are negotiated between the creator of the data and the consumer of the data. According to the results of the negotiation, the needed data is provided. According to [14], the negotiation module is inserted between the data consumer and the data producer. The goal of meditated the negotiation module is to preserve the privacy of data producers as well as use the producers' data to offers application services. A set of privacy policies controls the module under IoT system management.

The process of negotiation in IoT is to form an arrangement or kind of contract between data consumer and data producer which allows data consumer uses the data producer. The process consists of two main parts: Negotiation between the data producer & the module and Negotiation between the data consumer & the module.

In the negotiation session between the producer and the module, the producer of the data interacts with the negotiation module to publish his/her information. The negotiation module, then, produces a set of PII which represents the information that will not be revealed. The data producer publishes his/her data to permit certain services. Then, the negotiation module maintains sensitive information and its quasi-identifier that might cause privacy vulnerabilities. The negotiation module then shows the data producers how to protect attributes of the sensitive information. Depending on the response of the data producer the negotiation module will act. If the data producer follows the guidance steps, the data is going to be anonymized before it is published.

The second part of the negotiation process is the Negotiation between the data consumer and the module. The negotiation must be done between the data consumer and the negotiation module. A request needs to be made by the data consumer to start a negation session between him/her and the negotiation module. The negotiation module validates the request based on predefined privacy policies. If the request is invalid, the data consumer is notified that he/she cannot get information. Otherwise, a quick analysis is going to be performed to measure the sensitivity of the requested information. With sensitive information, data consumers should agree, and accordingly, a privacy preserving rule is established for the data consumer.

In [15], big data privacy via the hybrid cloud technique was introduced. The new technique includes a random one-to-one mapping function for images encryption. This function speeds up the process of replacement and dissemination and the private cloud stores just the mapping function parameters. It also describes the proposed architecture of the hybrid cloud. A private cloud contains a server that processes the data is passed through the private cloud and checks its sensitivity. If the data contains sensitive information, the data will be reprocessed to ensure that there is no sensitive data might be infiltrated out. Otherwise, the original information is going to be straight passed to the public cloud. The objectives of this design are to ensure the following overheads:

- The amount of data saved in a private cloud.
- The resulted overhead of interacting private cloud with public cloud.

- The delay is caused by interacting between the private cloud and the public cloud.

Also, the one-to-one mapping function is introduced and included in this scheme. The basic concept of the function is to map the original pixel value to a random value. The scheme splits the image into blocks and mixes the block with unpredictable positions. The proposed technique runs at the block level rather than pixel level to deliver faster computations. The mechanism offers secure storage as well as small overhead computations.

The services provided by cloud databases attract the attention for managing outsourced databases, but the privacy and security changes accompanying them drive the adoption to the cloud databases down. Encrypting the data before it is sent to the public databases servers has become a direct resolve to this issue. However, encrypting the databases makes responding to the queries more complicated. Also, privacy remains an issue even if the data encrypting in the cloud because the data will be fully under provider control.

Together homomorphic encryption schemes and order-preserving indexing schemes are incorporated in one scheme to facilitate querying encrypted databases. Moreover, the method used in the proposed technique can be implemented in current DBMSs. Paper [16] described and showed the architecture of the proposed technique. The concept of this architecture is simple and as it depicts contains from public database service cloud-connected to an enterprise. The data is encrypted by the enterprise. The query proxy unit is included within the enterprise domain and it is used to query or update the encrypted database, so it controls the communication between applications and encrypted database. The order-preserving scheme is used to achieve range queries on a homomorphically encrypted database.

In [17], a privacy-preserving protocol against similarity detection uses the concept of geometric transformation of some data is produced. This geometric transformation can be applied to vectors produced using cosine similarity to preserve privacy. The protocol ensures the confidentiality of the sensitive data and delivers secure clusters. Both Random scaling and Rotations are techniques used within the protocol to prevent similarity detection algorithms from revealing personal information.

The geo-tagged and location information in social media can be used to preserve the privacy of the users by delivering an awareness regarding the possibility of compromising media [18]. Also, the advantage of the mobility tracking feature of contemporary smartphones is considered to generate what is called a smart privacy zone. In this privacy zone, the users would be notified about media events. The basic concept is to emerge the time and the place into the location data stored in the media to decrease the amount of sensitive information related to the users.

## 3.2 Access Control Model Approaches in Big Data

Expensive computations at the users' side which are accompaniment by fine-grained access control policies are being an obstacle for cloud computing customers. This concern is being severe with access control of live data streams to the cloud. Considering combination between trigger and sliding window policies addresses this issue and allows the data owner to determine fine-grained policies linked to their data streams. In addition, the cloud will encrypt data on behalf of the owner as well as provide secure live processing over the data. Also, the cloud enforces the policies of access control, and the confidentiality of the data is protected at the same time as well as no possibility of unauthorized access. Achieving this comes possible with applying proxy-based attribute-based encryption and integrating the XACML framework for managing policies [19]. This approach delivers a high level of scalability while the number of policies and users is increasing.

According to [8] defining a suitable access control model for information dissemination systems is needed due to the lack of access control mechanisms. The user profile is the basis of an approach that can be used as a proper access control mechanism for information dissemination systems. This profile consists of two types of information. The user profile is categorized into user interests and credentials. Information objects of what the user is interested in acquiring are listed in the first category which is user interests. The second content of the user profile is the credentials that are used to classify the users to facilitate defining appropriate access control policies. Therefore, access control policies can be enforced according to user credentials contents. Accordingly, users will only acquire information interested to them as well as they are qualified by proper authorization. To ensure authenticity, the user might be required to provide some personal data certified by a third trusted party. The architecture of the proposed system contains a Filter module that is activated when new information objects are being added and it is responsible for selecting candidate users based on their interests to be notified. The results of the filtering step are passed to another module which is the Access Control module after it is activated. Then, the Access Control module validates whether the user is eligible to be notified or not based on its authorization state.

### 3.3 Encryption Approaches in Big Data

In [9] MapReduce, conventional encryption techniques and methods do not allow operations and computations over the encrypted intermediate data. This is considered a challenge for the MapReduce product. However, a fully homomorphic encryption scheme (FHE) is being a solution to address this challenge in MapReduce. This scheme facilitates computations and operations on encrypted data without not need to decrypt the data. Moreover, the FHE scheme also addresses the issue of protecting intermediate data in MapReduce. Therefore, the FHE scheme provides both data confidentiality and supports the computations of encrypted data. MapReduce first receives inputs formed pairs from key/value. Then, MapReduce uses Map to map similar keys. After the same keys are being mapped, they are passed to Reduce. The main task for Reduce is to join those values. During this procedure, the data of intermediate files are not secure. This indicates that data confidentiality is not ensured and supported. However, the FHY scheme overcomes these issues. Given a set of ciphertexts that encrypts elements of a set of plaintexts, the computations over plaintexts are possible by decrypting the results of computations that are running over a set of ciphertexts. If this concept is applied to intermediate data of MapReduce, the confidentiality of the data is protected as well as the ability to run operations over the data is possible.

In [20], Also, data anonymization is considered a potential solution for preserving privacy and protecting the data in MapReduce. Data anonymization approaches provide scalability, flexibility, and efficient protection of the privacy layer on the framework of MapReduce in the cloud. This approach is proposed as the filtered layer that checks the data before it is being accessed and processed. The layer proposed is on top of the MapReduce framework. The layer introduces a gateway for the clients to determine the requirements are needed for preserving the privacy of data based on the variety of privacy predefined models. After the privacy requirements are defined and specified, the proposed layer will introduce data anonymization mechanisms to MapReduce. The anonymized data is going to be kept because it will be reused when it is needed, so this will avoid the expensive cost of re-computations. Therefore, the data is dynamically updated and maintained.

In [10], MS2 is an approach that addresses the security issue in the cloud-based infrastructure and services. The approach provides a hybrid encryption technique that supports encryption and decryption in real-time in the cloud. Both cloud service types that are computing and storage are considered in the proposed approach. The approach supports data verification and data encryption during data rest at the cloud and data transit. The identity-based encryption technique is the basic concept of this approach. In identity-based encryption, the cryptographic key is derived by the user's identity such as a secret key when AES is the used crypto function. Securing the privacy provided by distinguishing sensitive data from non-sensitive data. The operations over the sensitive and critical data will be allowed after the second level when the user is verified and validated. Consequently, the user's identity is ensured to not be misused. On another hand, if the data is not critical and sensitive, the second verification step will be disregarded to enhance the performance of the cloud. From the identity of the user, the key is derived and generated. The generated key is used for encrypting and decrypting the set of data related to each user. The architecture involves encryption service and verification services. The encryption services depend on password-based encryption techniques and identity-based encryption techniques. Also, they offer selections of encryption functions as well as second-level verification services. In addition, the user has the choice of performing data encryption whether in their side of the cloud end. This increases the level of trust of the cloud as well as supports flexibility environment in the cloud. Also, transparency for data security is supported when the operations are performed over the data.

In [11], the proxy re-encryption (PRE) mechanism could not protect privacy in cloud computing and big data efficiently. Another approach was proposed to address the drawbacks of the (PRE) such as the conditional proxy re-encryption approach CPRE. In this CPRE mechanism, the data originator chooses a condition value to use its private key for encrypting the message and the re-encryption key is generated from this condition value and members' public key. If any member left the group, the condition value is going to be changed. Accordingly, the process of computing the condition value and encrypting the message will be altered. Therefore, a new member could not decrypt old information as well a member who left the group could not decrypt new messages. However, the size of the group will cause overhead computations when it becomes large or the group changes repeatedly. The efficient CPRE was proposed to address the changing of the group members. In E-CPRE, the re-encryption key does not need to be created and uploaded for all group members. The drawback of both CPRE and ECPRE mechanisms is that the current messages need to be encrypted using the new condition value. However, an enhanced version of the E-CPRE mechanism was introduced and called outsourcing CPRE (O-CPRE). The O-CPRE mechanism requires the message's originator to execute two main tasks when the group members are changed. First, the message's originator will compute the condition value changing key (CCK). Then, the message's originator should submit the CCK to the cloud. This will reduce the overhead caused by membership changing because the cloud storage will convert current ciphertexts using received CCK.

### 3.4 Authentication Approaches in Big Data

The concept of data storage in the cloud is to use a particular main server connected to multiple servers to store user data. Using recovery servers connected to the main server to retrieve the data in disaster cases is a mechanism based on authentication and authorization enforcement processes [21]. The main server is positioned and remotely connected to three back-ups severs. Proper encryption and decompression methods are used to decrypt and compress data during recovery and backup operations. The user interacts with the main server to store and retrieve data. To login to the main server, the main server first authenticates the user using his email as an id and his password and generates a secret key. Then the secret key is used for authorizing the user to another login interface. Also, the secret key is used for encryption and decryption security services used in the system involved multi-server.

According to [12], exposing users' sensitive information to untrusted cloud servers is one of the Multi-factor authentication (MFA) approaches, which is used for user validation in a cloud server, concerns, and challenges. MFA system, usually, uses two or more factors for validating the user. Physiological characteristic identification elements might be considered as one of the factors that validate the user such as a fingerprint. Untrusted cloud servers might misuse that sensitive information. Privacy-preserving multi-factor authentication system named MACA is proposed to address security issues related to the MFA systems. The main idea of MACA is to aggregate hybrid user behavior profiles that are considered as a second authentication factor with the first factor which is usually a user's password. Several categories of user behavior features are incorporated and considered in the generated profiles. Also, fully homomorphic encryption (FHE) and fuzzy hashing methods produced in literature are included in this approach to ensure the security of user-profiles and sensitive information. Moreover, during the authentication procedure, the MACA approach capably integrates big data features. To generate profiles for users, two main components are incorporated network-based features and host-based characteristics. Also, four main components are integrated and form the architecture of the proposed system. An open-source profile acquisition program (PAP) is located in the user's host, an authentication server (AS) responsible for verifying the user login to the cloud, a user profile database (UPDB) which contains users' data with privacy protection, and content server. When the user interacts with MACA, the acquisition program generates a user profile which is then hashed and encrypted using FHE. Then the AS entity receives the user ID and password. Then, AS interacts with UPDB to store the user profile database. Any login attempt is going to be evaluated by the AS by measuring the differences between the current and newly generated user profile. If the AuthResult succeeds and the user is authenticated, the content server ticket will be sent to the user.

## 4. Conclusion

The big data era has been already declared. Analyzing a massive amount of data cloud delivers better enterprises and advanced services. However, technical challenges need to be addressed and solved to deliver professional generating, collecting, and processing of data. The challenges include scale, heterogeneity, timeliness, visualization, data errors and incompleteness, privacy, access control, encryption, and authentication.

In this work, a brief introduction was given regarding big data. Also, the requirements of the big data mentioned include collection, storage, and processing phases. In addition, some big data benefits are supported by various examples such as healthcare, smart grid, traffic management, retail, and payments.

In this work, also, the security and privacy concerns and challenges are discussed for each of privacy, access control, encryption, and authentication. Moreover, some approaches and proposed techniques are given. Techniques and approaches address some security issues in big data including privacy, access control, encryption, and authentication.

## References

[1] Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In *2013 international conference on collaboration technologies and systems (CTS)* (pp. 42-47). IEEE.

[2] Jain, P., Gyanchandani, M., Khare, N., Singh, D. P., & Rajesh, L. (2017). A Survey on big data privacy using hadoop architecture. *International Journal of Computer Science and Network Security (IJCSNS)*, *17*(2), 148.

[3] Tene, O., & Polonetsky, J. (2012). Big data for all: Privacy and user control in the age of analytics. *Nw. J. Tech. & Intell. Prop.*, *11*, xxvii.

[4] Smith, M., Szongott, C., Henne, B., & Von Voigt, G. (2012, June). Big data privacy issues in public social media. In *2012 6th IEEE international conference on digital ecosystems and technologies (DEST)* (pp. 1-6). IEEE.

[5] Dong, C., Chen, L., & Wen, Z. (2013, November). When private set intersection meets big data: an efficient and scalable protocol. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security* (pp. 789-800).

[6] Patil, H. K., & Seshadri, R. (2014, June). Big data security and privacy issues in healthcare. In *2014 IEEE international congress on big data* (pp. 762-765). IEEE.

[7] Liu, W., Uluagac, A. S., & Beyah, R. (2014, April). MACA: A privacy-preserving multi-factor cloud authentication system utilizing big data. In *2014 IEEE Conference on*

*Computer Communications Workshops (INFOCOM WKSHPS)* (pp. 518-523). IEEE.

[8] Fang, W., Wen, X. Z., Zheng, Y., & Zhou, M. (2017). A survey of big data security and privacy preserving. *IETE Technical Review*, *34*(5), 544-560.

[9] Bertino, E., Ferrari, E., & Pitoura, E. (2001, April). An access control mechanism for large scale data dissemination systems. In *Proceedings Eleventh International Workshop on Research Issues in Data Engineering. Document Management for Data Intensive Business and Scientific Applications. RIDE 2001* (pp. 43-50). IEEE.

[10] Chen, X., & Huang, Q. (2013, May). The data protection of MapReduce using homomorphic encryption. In *2013 IEEE 4th International Conference on Software Engineering and Service Science* (pp. 419-421). IEEE.

[11] Raghuwanshi, D. S., & Rajagopalan, M. R. (2014, January). MS2: Practical data privacy and security framework for data at rest in cloud. In *2014 World Congress on Computer Applications and Information Systems (WCCAIS)* (pp. 1-8). IEEE.

[12] Son, J., Kim, D., Hussain, R., & Oh, H. (2014, April). Conditional proxy re-encryption for secure big data group sharing in cloud environment. In *2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (pp. 541-546). IEEE.

[13] Chakravorty, A., Wlodarczyk, T., & Rong, C. (2013, May). Privacy preserving data analytics for smart homes. In *2013 IEEE Security and Privacy Workshops* (pp. 23-27). IEEE.

[14] Dev, H., Sen, T., Basak, M., & Ali, M. E. (2012, November). An approach to protect the privacy of cloud data from data mining based attacks. In *2012 SC Companion: High Performance Computing, Networking Storage and Analysis* (pp. 1106-1115). IEEE.

[15] Ukil, A., Bandyopadhyay, S., Joseph, J., Banahatti, V., & Lodha, S. (2012, August). Negotiation-based privacy preservation scheme in internet of things platform. In *Proceedings of the First International Conference on Security of Internet of Things* (pp. 75-84).

[16] Huang, X., & Du, X. (2014, April). Achieving big data privacy via hybrid cloud. In *2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (pp. 512-517). IEEE.

[17] Kumar, A., Lee, H., & Singh, R. P. (2012, October). Efficient and secure Cloud storage for handling big data. In *2012 6th International Conference on New Trends in Information Science, Service Science and Data Mining (ISSDM2012)* (pp. 162-166). IEEE.

[18] Leontiadis, I., Önen, M., Molva, R., Chorley, M. J., & Colombo, G. B. (2013, September). Privacy preserving similarity detection for data analysis. In *2013 International Conference on Cloud and Green Computing* (pp. 547-552). IEEE.

[19] Liu, D., & Wang, S. (2012, October). Query encrypted databases practically. In *Proceedings of the 2012 ACM conference on Computer and communications security* (pp. 1049-1051).

[20] Dinh, T. T. A., & Datta, A. (2012). Stream on the sky: Outsourcing access control enforcement for stream data to the cloud. *arXiv preprint arXiv:1210.0660*.

[21] Zhang, X., Liu, C., Nepal, S., Dou, W., & Chen, J. (2012, November). Privacy-preserving layer over MapReduce on cloud. In *2012 Second International Conference on Cloud and Green Computing* (pp. 304-310). IEEE.

**Sultan Alotaibi** received t the B.Sc. degree in Computer Science from Umm Al-Qura University, the M.Sc. and Ph.D. degrees from University of North Texas, USA. He is currently an Assistant Professor with the Department of Information Systems, College of Computers and Information Systems, Umm Al-Qura University, Saudi Arabia. His research interest includes wireless communication networks, LTE, LTE-A, radio resource scheduling, 5G and data quality in wireless networks.