Journal of King Saud University – Computer and Information Sciences 34 (2022) 10217–10245

HOSTED BY

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Data provenance for cloud forensic investigations, security, challenges, solutions and future perspectives: A survey

Oludare Isaac Abiodun [a,*], Moatsum Alawida [b,*], Abiodun Esther Omolara [a], Abdulatif Alabdulatif [c]

[a] Department of Computer Science, University of Abuja, Gwagwalada, Nigeria
[b] Department of Computer Sciences, Abu Dhabi University, Abu Dhabi 59911, United Arab Emirates
[c] Department of Computer Science, College of Computer, Qassim University, Buraydah 52571, Saudi Arabia

ARTICLE INFO

Article history:
Received 1 September 2022
Revised 14 October 2022
Accepted 18 October 2022
Available online 25 October 2022

Keywords:
Cybersecurity
Cloud computing
Data provenance and challenges
Solutions
Security
Digital forensic

ABSTRACT

It is extremely difficult to track down the original source of sensitive data from a variety of sources in the cloud during transit and processing. For instance, data provenance, which records the origins of data, and the record of data usage, update and processing can be introduced to trace malicious vulnerabilities. Thus, data provenance process makes it easy to monitor the sources and causes of any problems in cloud computing. However, data provenance is one of the most prominent drawbacks in cloud storage. Despite many studies, a full assessment of data provenance in cloud forensics is still missing from the literature, especially in wireless sensor networks, blockchain, Internet of Things (IoT), security and privacy. Importantly, one of the major challenges in data provenance is "how to reduce the complexity of evidence." That is, ensuring volatile data is captured before being overwritten. Hence, this study presents a survey of recent data provenance problems in cloud computing, provenance taxonomy, and security issues. It also, discusses how volatile data can be captured before being overwritten and then helps identify current provenance limitations and future directions for further study. More also, it examined how data is collected as evidence for digital crime in a real-world scenario. Furthermore, future work in digital provenance for cloud forensics, wireless sensor network, IoT, and blockchain is recommended.

Contents

* Corresponding authors.
E-mail addresses: oludare.abiodun@uniabuja.edu.ng (O. Isaac Abiodun), moatsum.alawida@adu.ac.ae (M. Alawida).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2022.10.018

1. Introduction

Cloud computing has captivated a lot of scientific research interest because of its many benefits and promising future (Liu et al., 2021; Bayramusta & Nasir, 2016; Ahmed et al., 2012; Srinivasamurthy & Liu, 2010). In addition, its services' low cost has contributed to their increasing popularity among individuals, companies, and governments. Although cloud computing offers attractive service packages for the commercial sector, such as access to the Internet, accessible communication, storage opportunities, and electronic commerce (Alhomdy et al., 2021; Xue and Xin, 2016), nevertheless, its adoption is still slow mostly due to the privacy and security concerns (Al-Hujran et al., 2018; Akin et al., 2014). Cyberspace or cloud computing has dramatically increased computers' practical utility and peripherals. It is widely used in finance, business management, telecommunications, transportation, education, healthcare and other areas of our daily lives. It also enables users to communicate efficiently, share software, hardware, as well as data resources via network protocols.

More also, it fosters breakthrough advancements in industry and scientific research, as well as a better quality of life for people. The Internet of Things (IoT) is a network that connects all physical intelligent items, no matter how diverse, complicated, or dynamic, to the Internet—enabling communication between physical and cyber-worlds, which can be used in different fields to enhance intelligent services as well. Nevertheless, there has been a slew of computer network vulnerabilities and IoT security and privacy issues during the past several years due to cyber-insecurity or cyber-attacks (Abiodun et al., 2021a,b; Ding et al., 2019; Bertino, 2016; Yan et al., 2014). Managing identities, ensuring data integrity, detecting malicious activity, and restricting access to different types of data in a network and IoT systems are all complex tasks as more crime investigations involve strategic digital tracings in increasingly large numbers and complexity. Cloud computing has helped to reduce the cost of data storage. However, the current cloud computing approach has raised many concerns regarding security and privacy.

Forensics of digital devices focuses on recovering and preserving lost or destroyed data, whereas cyber security attempts to prevent data breaches or other cybercrimes from happening in the first place. Evidence of digitize crime must be identified, preserved, examined and documented in order to be considered admissible in digital forensics. This is being done in order to be able to provide evidence in the event of a criminal trial. Therefore, digital forensic (Arshad et al., 2022) focuses on evidence presentation on crime to identify the offender and reduce computer crimes as well as reduce widespread concerns about personal data insecurity.

Data provenance solves cyber-attack issues by keeping information on the origins of data, usage, modification, including the person that accessed the data, purpose, date and time—then making it easy to digitally find the source of any committed crime or problem. Provenance is the record of a document, file, or data that helps in knowing its origin and history. Therefore, the record can serve as evidence in any occasion, case or litigation. This record can also serve as evidence of either document, file, data, image or video creation, processing, compression, transmission, transcoding, update, manipulation or deletion in terms of time, date and the person involved. In a digital forensic investigation, the evidence of an attack can be used to demand justice in a criminal court. The validity and repeatability of scientific findings depend heavily on the provenance of research data. The capturing of provenance in a consistent, machine-actionable form becomes increasingly crucial as research infrastructures offer researchers more combined datasets and more integrated facilities for processing and publishing data. Therefore, this study focuses on efficacious data provenance-gathering mechanisms required for cloud forensics. The current state of provenance methodologies and state of the art for cloud provenance are examined in this study. It presents a survey of recent provenance problems in cloud computing, focusing on IoT, blockchain, sensor networks, and helps identify current cloud provenance limitations for further study.

The paper is structured as follows: Section 1 introduces cybersecurity issues in cyberspace. Section 2 highlighted the contributions of this paper. Section 3 gives an overview of cyber-attack challenges in cloud computing. Section 4 brings to light paradigms with respect to cyber-threat in cloud computing. Section 5 discusses strategies for combating computer crime. Section 6 explains digital forensics. Section 7 describes cloud computing. Section 8 dwells on provenance. Section 9 discusses data provenance. Section 10 explains data provenance in IoT. Section 11 discusses sensor networks in data provenance. Section 12 describes provenance collection requirements. Section 13 enumerated blockchain-based data provenance for the Internet of things. Section 14 discusses applications of data provenance. Section 15 explains challenges in data provenance in cloud forensics. Section 16 describes IoT data provenance challenges. Section 17 discusses modern solutions to cloud provenance. Section 18 presents data provenance research gap in cloud forensics. Section 19 lays out the future directions. Section 20 concludes the study.

2. Contributions to knowledge

This study explores variants of cybersecurity incidences and challenges in proving the integrity of data in the cloud, which bring about forensic investigations. The contributions of this paper include the following:

1. This article equips the reader abreast of existing problems of cybersecurity, cloud computing and emerging trends in data provenance for digital evidence of a crime.
2. It provided a taxonomy of data provenance and elaborated data collection as evidence for digital crime in a practical world.

3. The article also investigated challenges associated with data provenance. It offered opportunities for participants to air their views on the critical aspects of data provenance uses, limitations, challenges and possible solutions.
4. Moreover, the study assesses data provenance, compares performances, and discusses future perspectives.
5. It covered areas of cloud computing, data provenance, IoT, blockchain, and digital forensics that have impacted the socio-economic well-being of the people and,
6. Furthermore, it discussed the challenges the cloud computing industry faces concerning data provenance and how to solve them.

Therefore, this study's findings will serve as a guide for creating, developing, and implementing more data-provenance strategies that can counter cyber-attacks. An overview of the primary contributions of this study as well as the uniqueness of its methodology are presented in Fig. 1.

Fig. 1 depicts the research area of contribution which are cyber-attack, cloud computing, data provenance in IoT, data provenance in blockchain, and data provenance in sensor network system. Others include digital forensic, data provenance challenges, research gaps and future trends.

3. Cyberattack

A cyberattack is perpetrated on the Internet to disrupt, disable, destroy, or maliciously manipulate a computing environment and infrastructure by remote adversaries (Zargar et al., 2013). It can also be deployed to destroy the data's integrity, steal data, control information, for ransom or any other malicious intention (Huang et al., 2018). A cyberattack is the deliberate use of computer networks to launch attacks on computer systems and their peripherals, such as servers, networks, and underlying infrastructure. These attack techniques include hacking, malware, DoS, DDoS, flooding (Jensen et al., 2008), or advanced persistent threats (APT) (Alawida et al., 2022), and spam. For example, in a data flooding attack, the attacker could send a sizable number of link requests to the target computer until all the target computer's resources are depleted (Aneja et al., 2018; Prasad et al., 2012). In addition, flooding can refer to a series of concurrent but geographically dispersed attacks on a specific business's authentication protocol or network log-in.

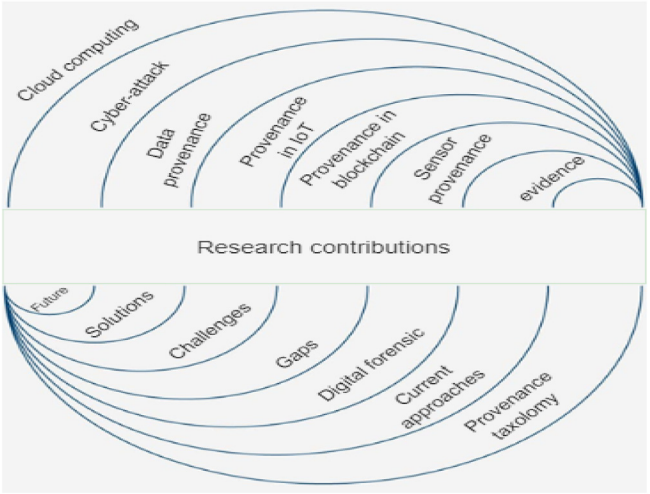

Fig. 1. Main contributions.

Similarly, the attack can be carried out by leveraging computer viruses or other means of malicious invasion and unauthorized access (Subramanya & Lakshminarasimhan, 2001; Zhang, 2013). Cyber-attacks have evolved into a type of asymmetrical warfare that is of great concern not only to computer scientists but also to the international community. Regardless, the main danger of a cyber-attack is the use of a computer as a tactic by a malicious adversary and a coercion multiplier (Huang et al., 2018; Woods, 2015).

One of the terrifying types of cyber-attack is the penetration of a military-sensitive infrastructure, which results in missile launches and destruction (Abbott, 2018). Other types of cyber-attacks target the financial and banking industries, the health sector, power infrastructure, and emergency services. Other scenarios proposed by researchers include unauthorized computer access, followed by data modification and theft. Errors, accidents, and natural disasters are all threats to computer and communication systems (Loch et al., 1992; Russell et al., 1991). Human operations, procedural errors, incorrect hardware, faulty software, dirty data problems, and electromechanical problems are all examples of errors. A cyberattack can originate from anywhere. An individual or a group can carry out the attack by employing one or more tactics, techniques, and procedures (TTPs).

Individuals who conduct cyberattacks are known as cybercriminals, hackers, bad actors, or threat actors. They can act alone, with other attackers, or as part of an organized criminal group. They attempt to identify vulnerabilities in computer systems and utilize them to further their objectives. The global cost of cyberattacks is expected to rise by 15 % per year, reaching more than $10 trillion (Cassim et al., 2020; Ventures, 2019). Ransomware attacks, which now cost businesses in the United States of America (USA) $20 billion per year, are a growing part of this cost. The average cost of a data breach in the USA is $3.8 million. Another concerning statistic is that after a successful breach, public companies lose an average of 8 % of their stock value.

4. Exposure to cyber-threat in cloud computing

Cloud computing has significantly improved data storage, processing, and transmission (Chen et al., 2021; As'habi et al., 2016; Bhisikar & Sahu, 2013). However, digital forensics in many types of cloud computing systems provides a daunting new challenge (Miranda Lopez et al., 2016; Grispos et al., 2012). Due to the complexity of cloud storage and the large number of users, conducting forensics in various cloud deployment models that may transcend geographical or legal boundaries has become a problem. Software services and their facilities require a virtualization environment across several different host computers connected to the Internet or an organization's internal network in a cloud computing model. For businesses and systems users, the cloud offers a common platform or service catalog within which virtualization can function. However, if the cloud is based on the Internet rather than the company's own network, it could place the company at greater risk of security and privacy breaches. Think about how data processing in a cloud computing system may occur across numerous locations and remain a mystery. There are commercial cloud service providers like Amazon Web Services, Microsoft Azure, and Google, as well as open-source cloud systems like Sun Open Cloud Platform. Cloud service delivery can be modeled in three ways:

(i) Software as a Service (SaaS) model, subscription-based or pay-*per*-use model for renting software.
(ii) Platform as a Service (PaaS) model, renting an application development environment from the customer.

(iii) Infrastructure as a Service (IaaS) model, on a subscription or pay-as-you-go basis, the hardware infrastructure is rented by the customer, and the service can be increased as required.

Thus, because of the nature of cloud computing services, an organization may not know where the data it is responsible for is located at any given time. Regional data storage options, such as European (EU) and United States (US) data storage, are available from several public cloud service providers. Because some cloud services replicate data, there may be more than one source for the same information. Cloud providers should be aware of where virtual machines are running, and they should keep a record of where they have been at any given point in time.

5. Combating computer crime

In recent years, criminal investigations now rely heavily on digital forensics. Information technology (IT) investigative and analysis methods are used to locate and analyze potential evidence. Crimes using computers and related technology crimes are usually broken down into three categories: the computer, which is either the target or a repository for information generated or used during the execution of the crime. This type of description is more typical, covering operating systems, data, software, and the investigation into actual computing devices that have been seized and analyzed. However, this analytical paradigm is no longer relevant when it comes to data distribution in the cloud computing environment.

Standards and processes for identifying, recovering, preserving, and analyzing computer-based evidence are becoming increasingly important as enterprises seek computer forensic services or investigations. An understanding of the computer forensic procedure, as well as an organization's legal responsibilities, is necessary to ensure the integrity and robustness of such investigations. Both the evidence collected and the procedures utilized in the gathering, preservation and analysis of digital data must demonstrate integrity in any computer forensic inquiry. If the inquiry uncovers criminal behavior, the admissibility of digital evidence in court is highly unlikely unless the integrity of the evidence is preserved.

Digital evidence can be challenging to get. Digital evidence can be more celestial, dynamic, and imaginative in a cloud computer system's virtual surroundings. Using cloud computing, data that would ordinarily be written to the operating system, such as registry entries or temporary Internet files, would reside or be saved in the virtual environment and, thus, will be deleted when the user quits the virtual environment. Due to virtualization's ability to sanitize resources, traditional analysis of remaining artifacts may be constrained. As a result, data stored on hard drives may be lost forever.

6. Digital forensics.

1. Digital Forensics.

The preservation, identification, extraction, and documentation of digital evidence that can be used as proof of a committed crime in court is known as digital forensics. Identification, Preservation, Analysis, Documentation, and Presentation are all steps in the digital forensics process. The categories of forensics include disk forensics, network forensics, wireless forensics, database forensics, malware forensics, email forensics, memory forensics, and others. Cases involving (a) fraud investigations, (b) intellectual property theft, (c) employment conflicts, and (d) industrial espionage can all be handled using digital forensic science. There are many challenges in digital forensics, including the rise in personal computers and widespread usage of the internet, easy access to hacking tools, and prosecution, which is challenging due to a lack of tangible proof. This investigation work is challenging because of the vast amount of storage space in terabytes. Any technological advancement necessitates a solution upgrade or modification. Categories and subcategories of digital forensics can be represented in Fig. 2.

Though there are many challenges of cloud forensic as shown in Fig. 2. However, this study focuses on the provenance aspect.

2. Models of the digital forensics process.

Models of the digital forensics process have been constructed, including at least eight unique processes and characteristics as follows:

(a) **Search authority** A search and/or seizure of information requires proper legal authorization.
(b) **Chain of custody** A traceable line of possession to prevent accusations of misconduct or evidence tampering—it is essential in legal settings to keep detailed records of who had access to and handled various pieces of evidence.
(c) **The ability to image and hash data** A precise copy of any artifact that may include digital evidence should be made and then hashed to confirm that it is a true copy.
(d) **Validated tool** The tools that have been proven to work should be discovered. In addition, there should be as much validation of forensics tools as possible to make sure they are accurate and trustworthy.
(e) **Analysis** The strategy is to analyze the situation. Investigative and analytical methods are used in forensic analysis to dissect and decipher retrieved physical and digital evidence.
(f) **Repeatability and reproducibility (quality assurance)** Forensic analysis should be performed and replicated with similar results by the same forensic expert or others.
(g) **Reporting** Providing a report on findings. The forensic analyst's analytical process and findings must be documented so that they can be used by other investigators.
(h) **Presentation.** The forensic analyst will give a presentation to a judge or other audience, outlining the case and the reasoning behind his or her findings. That is, provide the court with evidence and control chains.

These procedures must be performed or adapted to the cloud setting before digital forensic investigations can be conducted there. Many of them provide formidable difficulties. Part of the purpose of this study is to guide on conducting a forensic investigation using data extracted from the cloud. Cloud forensics is a similar field that we do not fully get into here but which is becoming increasingly important in the modern era. The cloud can be used to examine and analyze digital evidence (Herman et al., 2020; Martini & Choo, 2014). Digital forensic science's investigative process is shown in Table 1.

3. Evaluations of the digital evidence.

Evaluation of digital evidence involves admissibility, authenticity, completeness, reliability and believability. Digital evidence is practical proof of the reality saved or communicated in digital form. It can also be defined as "binary information that is saved or received, and that is admissible in court" (Saleem, 2015; Guo et al., 2010). Furthermore, it could be defined as a computer or digital-based information that substantiates and disproves a notion of how a crime was committed or addresses important aspects of the crime, such as motive or an alibi (Casey, 2011). The following legal considerations for gathered evidence should be made in accordance with RFC 3227 guidance for evidence collection and archiving.

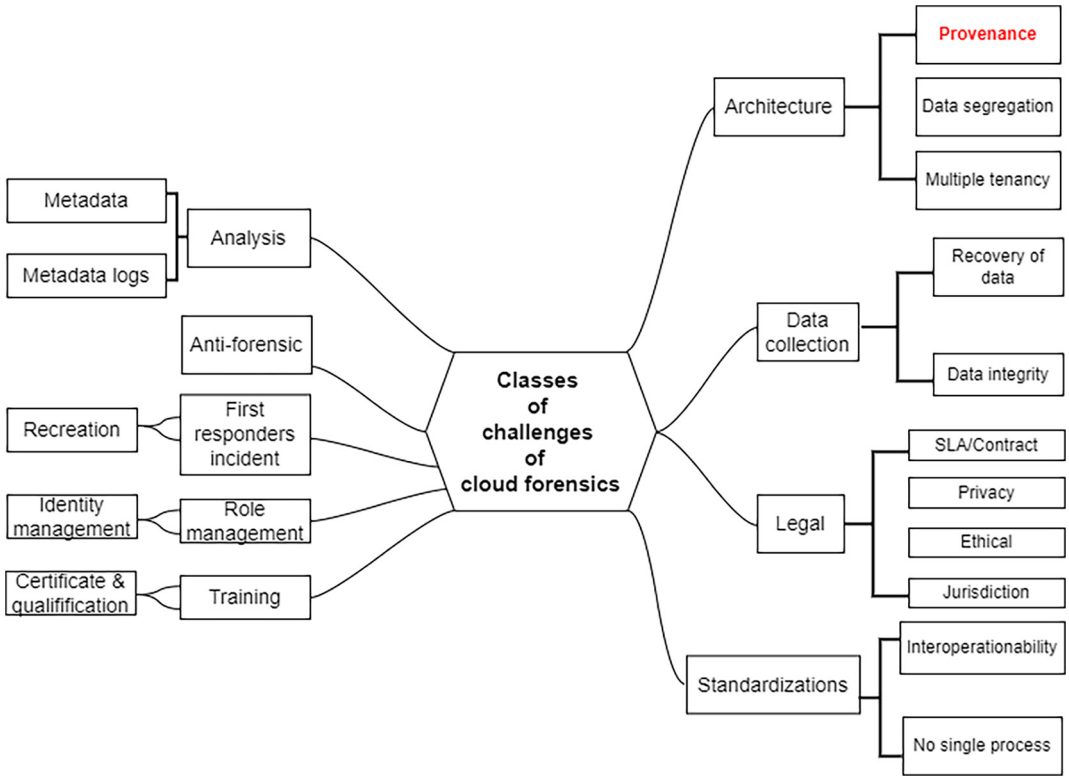

Fig. 2. Classes and subclasses of digital forensic.

Table 1
Digital Forensic Science's Investigative Process.

Identification	Preservation	Collection	Examination	Analysis	Presentation	Admissibility	Decision
Detection of crimes and incidents	Management of a case	Protection or conservation	Protection or conservation	Protection or Preservation	Record keeping	Presentation	–
Decision on the signature	Image innovation	Accepted approach	Tracking and tracing	Tracking and tracing	Witness from an expert	Compliance with constitution	–
Profile recognition	Chain or sequence of custody	Accepted software	Validation or verification methods techniques	Empirical, numerical or statistical analysis	Clarity of evidence	No doubt on evidence	–
Abnormal discovery	Synchronization with the time.	Acceptable hardware	Cleaning or filtering procedures	Links, protocols, algorithms used	Impact of the mission	Testimony from expert	–
Comments, concerns, complaints	Notification	Legitimate authority	Matching patterns	Analysis of data	Recommendations for defenses	Recommendations for defenses	–
System surveillance		Compressed without loss	uncovered data sources	Events, timeline, history	Interpretations of data	Interpretations of data	–
Analysis of audit		Survey, sampling	Hidden data extraction	Link		Proving of case	–
		Reduction of data		Special			–
		Recovery techniques		Artificial intelligence			–

Admissible. Digital evidence is admissible if it supports a claim made in court, was unchanged during the digital forensics process, and if the findings are credible, legitimate, and subject to peer review. In addition, evidence is legitimately admissible in court when it is offered to prove the facts of a case; and does not contravene the law or other legal statutes. Therefore, before being presented to a court, the evidence must abide by a number of legal requirements. (Cosic & Cosic, 2012).

Authentic. There are two requirements that must be fulfilled for evidence authentication. Firstly, electronic evidence needs to be legally gathered with the help of written authorization from the investigative authorities. Then it needs to be independently verified by computer science and IT professionals. The evidence is invalid if neither of those two requirements is met. Hence, there must be a mechanism to connect the evidence to the incident.

Complete. The evidence must convey the entire narrative, not simply-one viewpoint or part. Courts rarely offer assistance on where the balance lies. Hence, it is up to lawyers and investigators to make judgments based on challenging problems like;

What information is genuinely necessary to create a case's factual framework? Can the "full" evidence that has been provided be adequately authenticated? Where is the balance between too broad and too restricted in terms of privacy rights and how far back should a timeline go? Can a jury fairly consider incomplete evidence and find a defendant guilty? (i.e., would an appeal uphold the conviction?). What does all of this actually mean in terms of digital forensics examiners and lawyers?

(i) Printouts were popular as the most often used type of social media evidence in 2015. Even though software claiming to preserve posts and communications in a more natural state had been around for a while. The proof must first "accurately replicate the content and image of a certain webpage on the computer," and then it must be verified that it was actually written by the intended author.
(ii) Obtaining a "witness with expertise," such as an account's owner, to attest to a post's legitimacy is not the only way to verify text and images: Authenticating the evidence may be self-incriminating if the defendant is the account owner. The content could have been uploaded, altered, or removed by a third party with access to the account. The "webpage" or account could potentially be faked by an unidentified third party. Social networking is, by nature transient, and several mobile apps and features are built to automatically remove content after a predetermined amount of time.

A screenshot or even a photocopy may be the only way to prove that a tweet, a picture, or an Instagram Story existed at some point in time, despite all efforts to preserve the evidence. That is, such as turning a device into Airplane Mode or serving a platform service provider with a preservation order. It is also true, though, that screenshots and photocopies are fallible fixes. Screenshots are modifiable. Pages of printed text messages might not be entirely relevant, and poorly reproduced photographic evidence can be tough to evaluate.

(iii) Printouts and screenshots do not provide the entire scope or context of multimedia that may be included in a tweet, post, or text message's body. A screenshot or photocopy did not account for edited or other types of manipulated multimedia evidence until artificial intelligence made it possible to create "deepfake" pictures and videos. By comparing a piece of evidence's hash to others in a database like Project VIC, one can verify the authenticity of some evidence, including video proof of child sexual assault.
(iv) However, not every crime has databases of hashed unlawful photographs, and the technique is ineffective for newly created content. Notwithstanding, other factors are at play when determining the veracity of social media postings, messages, photographs, and videos. Also relevant is who posted it. Without those details, a reasonable juror could not determine the evidence's authorship. Therefore, courts have typically disregarded evidence that is not supported by a variety of other pieces of evidence.
(v) That is, alternatively, "evaluated as a whole with all the specific particular details taken into combination." A reasonable jury would not be able to determine who wrote the evidence without those specifics. Prosecutors and investigators frequently have a limited understanding of what "specific particular details" should be regarded as corroborant, even when there is more evidence than ever available. That is, partly due to the time and effort required to corroborate the evidence. However, using digital proof to completely substantiate strong victim statements is a potent strategy. A case is made weaker overall when one piece of evidence is not corroborated, increasing the likelihood that a court will reject it.

Reliable. The approach used to gather the evidence must be wholly reliable and authentic. The judiciary has unwavering faith in software's capacity to produce trustworthy evidence. This "Presumption of Reliability" is generally well-established in American law. For instance, the court stated that digital evidence provided a "prima facie aura of dependability" in Olympic Ins. Co. versus H. D. Harrison, Inc (Van Buskirk, & Liu, 2006; Fromholz, 1977; Olympic Ins. Co. v. H. D. Harrison, (1969). Rebutting the presumption of reliability is challenging. Most courts' judges will faithfully uphold the presumption of reliability until specific evidence is given to establish that the specific code in question has demonstrable flaws that are directly relevant to the evidence being offered up for admission. However, a party cannot audit code to evaluate its quality because the majority of it is closed source and closely guarded. However, source code audits may be the most effective method for finding flaws.

Believable. A court must find the digital evidence easily credible, non-complex, but understandable. As a result, the court can pronounce judgment in favour of the attack victim based on believable evidence. Then, the court can, at the same time pronounce judgment against the attacker based on the believable evidence.

7. Cloud computing

In recent times, cloud computing has grown in popularity as a method of data processing and storage (Chakraborti et al., 2022; Sohal & Sharma, 2022; Wang & Zhang, 2020; Langmead & Nellore, 2018). Cloud forensics is where network forensics and cloud computing converge. Its main objectives are to look into cybercrimes, ensure that regulations are followed, and look into user policy violations and data breaches. To do this, forensic analysts, cloud service administrators, and cloud service providers (CSP) need tools for tracking, examining, and analyzing data access patterns within the cloud environment (Jangjou & Sohrabi, 2022; Raghavendra et al., 2022; Kebande & Venter, 2018; Ko et al., 2011). In most cases, the physical location of cloud servers is unknown, which indicates that cloud forensics no longer accepts conventional seize and capture techniques (Manral et al., 2019; Trenwith & Venter, 2014). Notwithstanding, tracking data in the cloud regarding users, modifications, updates, and movement remains necessary. Provenance data is one method for accomplishing it.

(i) Cloud Dynamism.

In this section, we present a taxonomy of cloud infrastructures, the details of which have been identified in previous and current works (Jain, & Kumar, 2014; Oliveira et al., 2010; Hoefer & Karagiannis, 2010), and then discuss its dynamic nature and the challenges it poses for provenance collection. Cloud computing, compared to conventional customized implementations and solutions where information technology (IT) infrastructures are established and administered locally, provides systematic and standardized solutions by leveraging or utilizing centralized enabling services and distributed utilization of information systems. The following features distinguish it: resource pooling in which clients share cloud systems, quick flexibility in which service abilities size up then down in response to customer demand, ad-hoc self-service in which they could use services almost instantly, measured service in which they pay only for services utilized, more also wide network connection in which services are

accessible via the Internet and further connected through the use of heterogeneous client platforms. According to the design of IT resources, the cloud architecture can be shared into three main strata, that is, (i) physical, (ii) virtual, and (iii) application layers, as presented in Fig. 3.

Physical resources produce information about their reporting on status, security, and incidents. The generated data helps with incident investigation and security monitoring. Cloud providers and forensic investigation teams are the primary parties interested in this data. The physical layer includes the three basic storage, network, and server sub-layers of physical resources. That is, to support the virtual layer and cater to cloud users' demands, these physical resources can be properly interwoven and shared. Virtual resources produce information about their status, security, and incident reporting. However, they also generate utilization data that is utilized in the billing of IaaS cloud users. Cloud service providers, auditors, forensic investigators, and IaaS cloud users are the main parties interested in this data.

The application resources, on the other hand, compile data regarding their status, security, and incident reporting. When users use PaaS and SaaS clouds, they also generate usage data that is used to bill clients. The primary parties interested in this data are cloud providers, cloud customers, forensic investigators, and auditors. The application layer interfaces with cloud services in a variety of ways. The system layer is an additional layer that connects computer operating systems, servers, and storage to the physical and virtual levels (OS). Horizontal or vertical communication occurs between layers (Alnabelsi et al., 2022; Abbadi, 2011). When cloud resources communicate with one another within a layer, such as peer-to-peer file replication, it is known as horizontal communication. In order to support the three fundamental cloud computing service models IaaS, or infrastructure-as-a-service (Gauttam et al., 2022); PaaS, or platform-as-a-service (Zhang et al., 2022); and SaaS, or software-as-a-service vertical communication is required (Lee & Brink, 2020; Mell & Grance, 2011; Birk & Wegener, 2011). Iaas provides cloud users access to virtual computers that the Cloud Service Provider (CSP) has made available. They install their preferred operating system and applications, as well as monitor and continually manage their data on these machines.

PaaS gives cloud users access to digital development environments, which may be deployed and set up as programs that function in a particular environment. These environments include libraries and the OS of the CSP. When using SaaS, cloud users use application services offered by the CSP using client-side interfaces accessed through an API (API). Such programs can be run immediately from a web browser without needing any downloads or installations (Quick & Choo, 2013). However, the dynamic nature of cloud resources allows for the hosting of virtual resources across several physical locations. Additionally, a single application may be hosted across numerous physical servers, with policies strictly regulating and managing data transit between them (Jangjou, & Sohrabi, 2022; US-CERT, 2011). Scalability, multi-tenancy, high availability, resource consolidation, and resilience are all potential benefits of such a dynamic nature (Zhang et al., 2010; Armbrust et al., 2010). Conversely, this raises cloud security, privacy, lawful, auditing or monitoring, and logging concerns (Cook et al., 2018; Ruan et al., 2011), which must be addressed in order to help enable provenance.

(ii) Cloud computing network.

Cloud networking is an IT architecture where some or all of an organization's resources are stored in a private or public cloud platform, maintained in-house or by a service provider. Network resources and capabilities are stored and made immediately accessible on demand. Businesses today use the cloud to boost scale, create differentiation, drive agility, and expedite time-to-market. Again, for modern corporations, the cloud model has evolved into the de facto method for developing and delivering applications. Additionally, the way businesses handle their expanding infrastructure requirements, redundancy plans and regional expansions

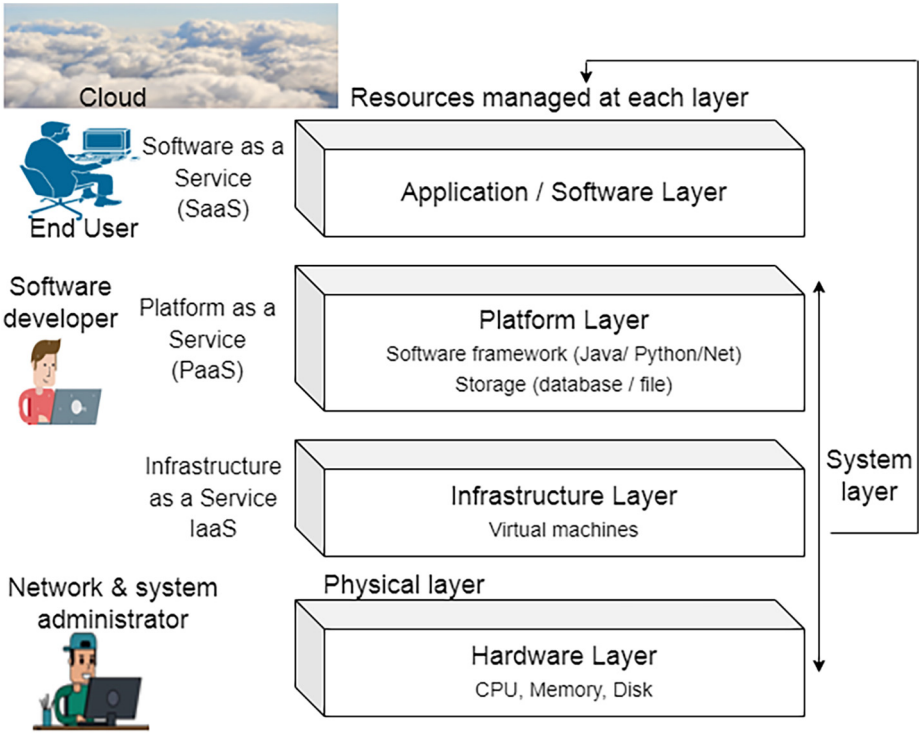

Fig. 3. Cloud computing architecture.

has been significantly influenced by cloud networking. Many businesses utilize different clouds from various cloud service providers and pursue a multi-data center strategy (CSP). A cloud computing network architecture housed in a private cloud platform is illustrated in Fig. 4.

A network diagram between ownCloud and its users is shown in Fig. 4. ownCloud is a self-hosted free software or open-source file synchronization and sharing server. In reality, the ownCloud architecture is set up on a central server machine and assigned a default IP address. Then, various nodes or clients are introduced into a network, including desktop personal computers, laptops, and mobile phones. The clients can use an IP address to access the ownCloud interface. Through a web-enabled interface, any client can upload and access their data to ownCloud. All data access events are tracked and logged by ownCloud for later auditing and analysis using well-known programs like Splunk. Authorized users can enable and disable features, set policies, generate backups, and manage users through the server's secure web interface, which administrators utilize to manage all of ownCloud's resources.

8. Provenance

Finding the data's origin and transforming it is significant and can be a difficult task as a result of the enormous development in data. It can be challenging to determine the source of a data object in many applications, including cloud computing, databases, and social media networks. In cloud computing, the original source of a shared data object is crucial for making a choice. A process to identify and document the source of data products is referred to as "data provenance." Both users and publishers of the data product have long acknowledged the significance of the data source. Scientists and researchers can use the original source of shared data items in the cloud to establish who is actually using these data and who is the genuine owner of these data products. Similar strict data quality checks are necessary for medical research because mistakes can endanger people's health.

A sort of metadata known as provenance, also known as history of records, pedigree, or lineage, ensures tracking of the steps for data derivation and provides fresh information for thought. Provenance can typically be described in various ways, including resource description frameworks (RDF), graphs, and other methods, depending on the domain type. Its provenance is the history of a data item that includes records and the passage of the data item via numerous owners (Sakka et al., 2012). Provenance is not limited to a single domain or field but has been addressed in a variety of application areas such as semantic web (Shvaiko and Euzenat; Moreau, 2010), scientific workflows and e-sciences (Cruz et al., 2009; Davidson & Freire, 2008), databases (Bell et al., 2013; Cheney et al., 2009a,b), curated databases (Buneman et al., 2006), services (Hastings et al., 2011), and file systems (Gehani et al., 2009). An in-depth description of the provenance lifecycle's main features is depicted in Fig. 5.

Three steps make up the provenance lifecycle: provenance storage, provenance query and analysis, and provenance collecting. Each phase has different requirements and difficulties depending on the domain. Considerations such as the scope, objectives, application domains, lifespan, and system assumptions are crucial.

Provenance is the record on a document or file that helps in determining its origin and history. This record can serve as evidence on any occasion, case or litigation. This record can serve as evidence of either document, file data, image or video creation, processing, compression, transmission, transcoding, update, and manipulation in terms of time, date and the person involved. A data object's provenance is described as metadata that contains details about its beginnings and past (Hu et al., 2020; Muniswamy-Reddy, & Seltzer, 2010). In addition, it provides information about the individuals, groups, and processes involved in producing a data object. Data provenance in cloud networks offers a way to track illegal activities. Since data stored in the cloud is disseminated broadly and covertly, provenance is crucial for cloud forensics investigators. However, the dynamic and distributed nature of the cloud makes obtaining provenance data challenging. These limitations give cybercriminals many opportunities to hide proof of their illegal operations.

Current cloud systems do not capture the whole provenance of digital objects, which limits cloud forensics attempts (Santra et al., 2018; Trenwith & Venter, 2014; Abbadi & Lyle, 2011). Therefore, further research is required to enable data provenance and give

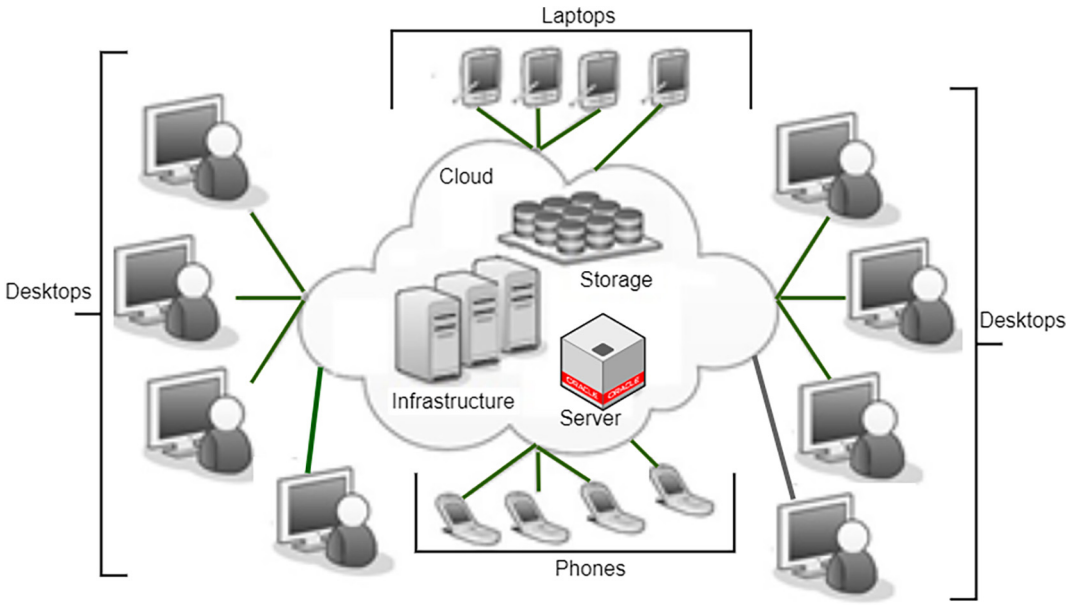

Fig. 4. Cloud computing network.

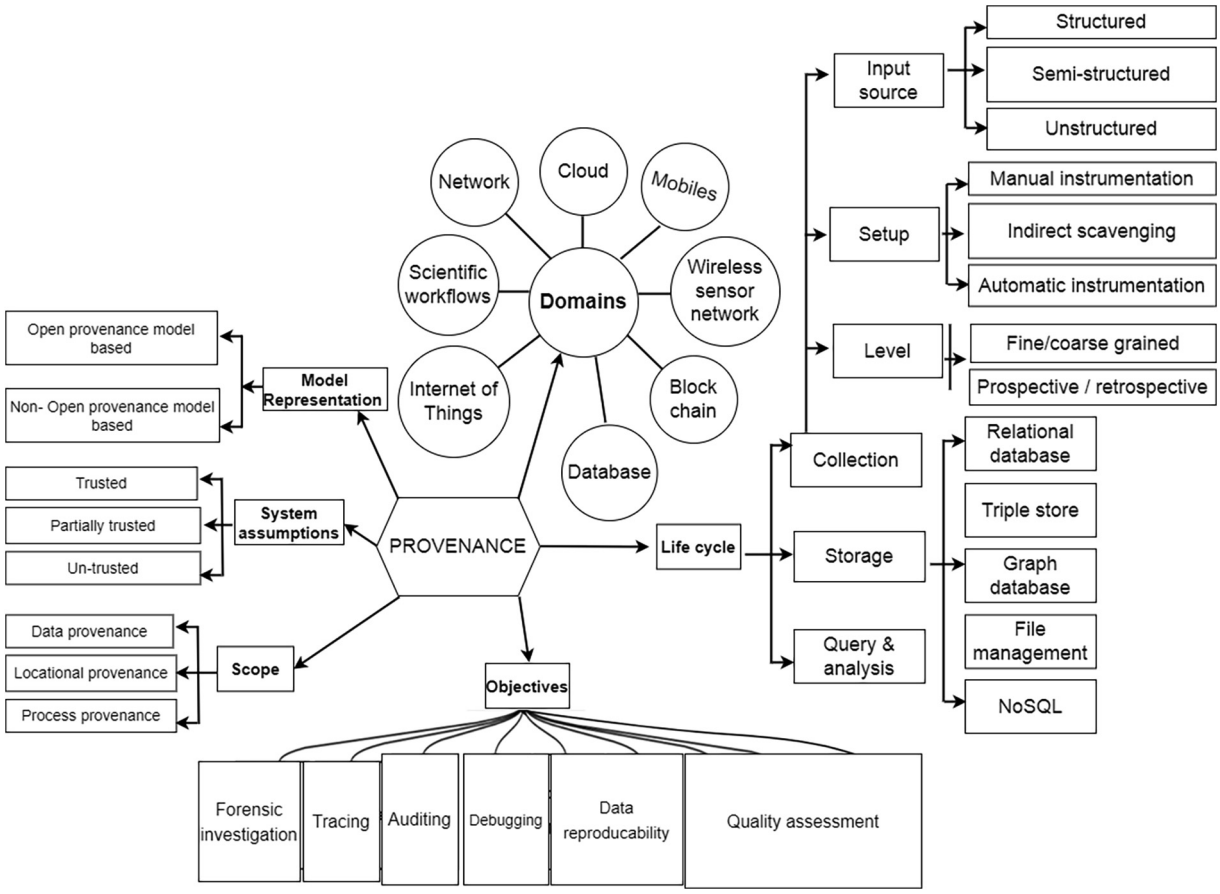

Fig. 5. General overview of provenance.

forensic investigators access to the precise location of cloud data and a more in-depth layout. This paper contributes in two ways: (a) it gives an overview of current provenance collection techniques in the context of cloud computing, along with their advantages and disadvantages, and (b) it identifies potential directions for future research to enhance data provenance in cloud forensics.

Ram et al. (2006) delved into the semantics or meaning of data provenance. They came up with a generic model they called the W7 model, which depicts data provenance as a set of seven interconnected elements: "What," "When," "Where," "How," "Who," "Which," and "Why." Each of these components has the potential as a source of provenance evidence. The W7 provenance model is a flexible, generic data provenance model designed to represent domains or applications with unique provenance needs. The provenance collection semantics are determined by the specific use cases of each application. Provenance in the W7 model is defined as an *n*-tuple, P = (What, When, Where, How, Who, Which, Why), where "What" denotes the sequence of events affecting the data object, "When" denotes the set of all timestamps related to the events, "Where" denotes the set of all locations of the events, "How" denotes the set of all actions preceding the events, "Who" denotes the set of all. An analysis of some techniques on data provenance is summarized in Table 2.

(A) Characteristics requirement for systematic management and capture of provenance.

Several attractive characteristics are required for systematic capturing and managing provenance:

i. It should have multi-level details of provenance information for easy understanding (Herschel et al., 2017; Cheah et al., 2013).
ii. It should have a provision for user annotations (Braun et al., 2008).
iii. It should have interoperability; that is, it should possess a characteristic to work with systems (Hasan et al., 2009; Cheney et al., 2009a,b).
iv. It should provide security assurances at a low cost for a provenance storage system that supports pruning (Zafar et al., 2017; Braun et al., 2006).
v. Privacy confidentiality query structure (Asghar et al., 2011)
vi. Provenance representation interpretation (Müller et al., 2018; Seltzer et al., 2005)
vii. Provenance capturing ought to be precise (Chan et al., 2019) without any argument (Cheah et al., 2013).
viii. Provenance capturing should have low computation overhead (Pasquier et al., 2017).

(B) Overview of provenance.

(i) Granularity Level: The amount of data that provenance is able to record depends on the level of granularity. In the context of scientific procedures, the granularity level is calibrated as both retroactive provenance and prospective provenance. In contrast, it is agreed upon as both coarse-grained and fine-grained provenance in the context of databases.
(ii) Provenance can be either prospective or retrospective. Prospective provenance records computation specifications that describe the procedures that will be followed to gener-

Table 2
Summary of Analysis of Data Provenance Techniques.

Scheme/System	Application domain/platform	Provenance semantics	Data granularity	Coarse/fine grained	Data status	Provenance computing	Storage strategy
Zero-information loss graph database (ZILGDB) (Rani et al., 2022)	Social media platform	Who, What, Where How, Why	Tuple	Both coarse and fine grained	Certain	Both	Tightly coupled
LineageChain (Ruan et al., 2021)	Blockchain applications	How, Where, What	Tuple	Fine grained	Certain	Lazy	Loosely coupled
System for provenance and data (Panda) (Patel et al., 2020)	Workflow	What, Where How, Why	Tuple	Both coarse and fine grained	Certain	Both	Tightly coupled
LineageChain (Ruan et al., 2019)	Blockchain applications	How, Where, What	Tuple	Fine grained	Certain	Lazy	Loosely coupled
ML Model (Buneman, & Tan, 2019)	Cloud/MXNet/scikit-learn/SparkML	How, Where, What	Tuple	Both coarse and fine grained	Certain	Lazy	Loosely coupled
Atmolytics visual reporting tool (AVRT) (Xu et al., 2018)	Information Visualisation of User Activities /Collaborative data sharing systems (CDSSs)	How, Where, What	Tuple	Both coarse and fine grained	Certain	Lazy	Loosely coupled
ProvChain (Liang et al., 2017)	Cloud/Block chain	How, Where, What	Tuple	Both coarse and fine grained	Certain	Lazy	Loosely coupled
Application Tracer (Badharudheen et al., 2014)	Basic UML design diagrams of the application	What, How Where	All	Fine grained	Certain	Both	Tightly coupled
S2Logger (Ko et al., 2011)	End-to-end data tracking	What, How Where	All	Fine grained	Certain	Both	Tightly coupled
System for provenance and data (Panda) Abbadi, & Lyle, (2011).	Workflow	What, Where How, Why	Tuple	Both coarse and fine grained	Certain	Both	Tightly coupled
Provenance querying language (ProQL) (Karvounarakis et al., 2010)	Collaborative data sharing systems (CDSSs)	How, Where, What	Tuple	Fine grained	Certain	Lazy	Loosely coupled
System for provenance and data (Panda) (Ikeda, & Widom, 2010)	Workflow	What, Where How, Why	Tuple	Both coarse and fine grained	Certain	Both	Tightly coupled
Efficient querying and maintenance of network provenance at internet-scale (ExSPAN) (Zhou et al., 2010)	Networks	What, How Where	All	Fine grained	Certain	Both	Tightly coupled
Managing Fine-Grained Provenance on Data Streams (Ariadne) (Tan, 2007)	Networks	How, Where, What	Tuple	Fine grained	Certain	Both	Tightly coupled
Uncertainty and lineage database (ULDB) (Agrawal et al., 2006)	Probabilistic DB	How, Where, What	Attribute	Fine grained	Uncertain	Eager	Loosely coupled
Provenance management in curated database (PMCD) (Buneman et al., 2006)	Curated DB	How, Where, What	Attribute	Fine grained	Certain	Eager	Loosely coupled
Provenance to support active conceptual modeling (PROMS) (Ram, & Liu, 2006)	All	All	All	Both coarse and fine grained	Certain	Lazy	Decoupled
Lineage tracing for general data warehouse (LTDW), (Cui and Widom, 2003)	Data Warehouse	How, Where, What	Tuple	Fine grained	Certain	Lazy	Tightly coupled
Efficient Stream Provenance via Operator Instrumentation (ESvOI) (Codd, 1970).	Networks	What, Where	Tuple	Fine grained	Certain	Both	Tightly coupled

ate outcomes. When it comes to retrospective provenance, it keeps track of all the actions conducted, including the environmental conditions that led to the outcomes. In addition, retrospective provenance contains more specific details about how the process was carried out (Ghoshal, & Plale, 2013; Freire et al., 2008).

(iii) Fine-grained and coarse-grained data: Databases can store both fine- and coarse-grained data; the former is used to record events at the tuple level, while the latter is used to record events at the relational or table level. The transformation process for each resultant data tuple or dataset is included in the fine-grained provenance, together with details on all input data tuples or packets. Fine-grained provenance can therefore help to ensure that results can be replicated. However, information at the process or view level is recorded in a coarse-grained provenance. However, the repeatability of internal tuple updates and delayed insertions cannot be ensured by coarse-grained provenance (Triantafyllou et al., 2021; Rupprecht et al., 2020). Furthermore, while coarse-grained provenance is more condensed, fine-grained provenance is more detailed (Cummings et al., 2009).

(iv) Provenance storage: The two portions of the discussion on evidence storage. The coupling and dissociation of provenance storage from original data is the first thing to consider. The storage model is the second aspect to take into account.

1. Coupling and decoupling of provenance storage with original data: When storing the provenance, it is up for debate whether to keep the original or raw data together or separately (Silva et al., 2016; Bates et al., 2013). That is, each has benefits and drawbacks. By maintaining consistency between tightly tied provenance and the data, protecting provenance during backup, restoration, and copies, and other factors, managing provenance in the storage system, for instance, may be advantageous to the user. On the other hand, we must ensure the storage system is reliable if records are kept in the exact location as actual documents.

Common data and provenance management policies ensure that provenance information is accessible to every-one with access to the data. Sometimes, implementing security measures for provenance data calls for different tactics than those used for data. For example, one can separate access control policies for data and provenance by decoupling provenance storage. Additionally, the storage cost brought on by the exponential growth in provenance data presents performance and scalability issues and may have a negative effect on query performance (Glavic, 2012; Pasquier et al., 2018).

2. Storage Models: Various storage choices, including file systems (Mothukuri et al., 2021), graph databases (Woodman et al., 2017), and relational databases, are accessible for the storage of provenance information (Vicknair et al., 2010). It also incorporates triple storage and NoSQL (Kashliev, 2020; Wylot et al., 2017). However, the challenging parts of storage models are efficient query support, storage size, and inference support. Original data files in file-system storage are where provenance is kept. Spatial Data Transfer Standard (SDTS) and Video Embedding of Information for Lineage (VEIL) (Gehani, & Lindqvist, 2007; Simmhan et al., 2005a,b). Examples of file-system-based provenance storing include the Flexible Image Transport (FITS) format used in astronomy. A FITS file's header carries provenance information, and the header size increases as the file's provenance does. The file-system storage model's drawback is that it only offers a few provenances search options.

(v) Provenance Query and Analysis: Without the assistance of query and analysis, the collection and preservation of provenance is useless. Two categories of provenance inquiries exist:

(i) Attribute Lookup: Look for provenance metadata that satisfies a predetermined set of requirements.

(ii) Transitive Closure of Ancestry, Genetics, or Descendancy: Search or Query for data objects by stating provenance constraints.

Models of storage that demand that users formulate queries in languages like SQL (Gadelha et al., 2011; Barga, & Digiampietri, 2008). Querying provenance is crucial to Prolog (Bowers et al., 2008) and SPARQL (Golbeck & Hendler, 2008). The query's performance is a critical issue and several techniques have been utilized to minimize query turnaround time (Karvounarakis et al., 2010). Size performance and provenance queries are tightly related. As provenance rises, query performance suffers. The performance of queries can be enhanced using provenance trimming and indexing techniques. Schema design improvements enhance query performance in relational databases. The inferencing capabilities of RDF triple stores are noteworthy for provenance storing and analysis. Regarding supporting provenance queries, relational databases perform better than non-relational RDF triple stores. Contrarily, relational databases do not support inferencing, which led to the creation of relational RDF stores to facilitate sophisticated provenance queries and inferencing. A typical example is the relational RDF triple storage RDFProv.

Graph-based storage models can provide provenance inquiries based on graph analysis (Neo4j). Transitive closure queries are computationally expensive for large graph sizes. Cycles on a graph can be eliminated to improve graph analysis. A graph and a relational database are compared from the perspective of provenance in a study by Vicknair et al. (2010), considering elements like API and Query Language simplicity, agility, flexibility, and security. While attribute lookup searches perform well in relational databases, transitive closure queries perform better in graph databases. The effectiveness of retrieval queries in graph databases is intrinsically tied to the indexing method.

When analyzing query speed, secure provenance must be considered along with implementation primitives used to guarantee security. While signatures are frequently employed to provide a traceable chain of custody, cryptographic processes negatively impact query performance. Gehani & Lindqvist (2007), however, presented the idea of "Lazy Trust Establishment," which validates only the part of the lineage graph required to answer the provenance question as opposed to the full network, in order to increase the performance of provenance analysis.

At the levels of the operating system, process, and workflow, provenance management can be implemented. Each technique has benefits and drawbacks. At the kernel or user level, provenance records data and data-process dependencies in the operating system collection. It automatically creates comprehensive metadata without altering the current process or interfering with the application's infrastructure. On the other hand, provenance awareness at the process and workflow levels calls for automatic/dynamic instrumentation for provenance collection. The taxonomy of provenance is illustrated in Fig. 6.

(C) Provenance Inference (PI).

Provenance inference (PI) is the method used by static code analysis to draw inferences from any data or reasoning. The subject of PI has not been extensively studied or stressed, even though it offers a wide range of applications. This article's main objective is to provide a detailed overview of the body of research on provenance inference methods. Provenance inference taxonomy is represented in Fig. 7.

(D) Artificial Intelligent Systems.

With the growing use of AI-based systems to support human decision-making, data provenance—a record that details data sources and processing-holds great promise. Transparency, equity, accountability, and interpretability are the four cornerstones of responsible AI, which aims to prevent the devastating effects that biased AI systems can produce. An overview of how data provenance for responsible artificial intelligence systems can help create data provenance is shown in the current study by Werder et al. (2022). Then, how might biases caused by data origins and preprocessing be mitigated to create ethical AI-based systems? They wrapped up with a research agenda detailing potential future study topics.

A graph-based data architecture called the resource description framework (RDF) is utilized for online information sharing (Sikos & Philp, 2020). RDF has been standardized by the World Wide Web Consortium (W3C). The data model strongly emphasizes using Uniform Resource Identifiers (URIs) to identify every model component in a globally unique way. The combination of terms from many schemata in a single RDF graph is made possible by the globally unique identification of schema elements (referred to as vocabulary words in the RDF context), essentially making RDF a

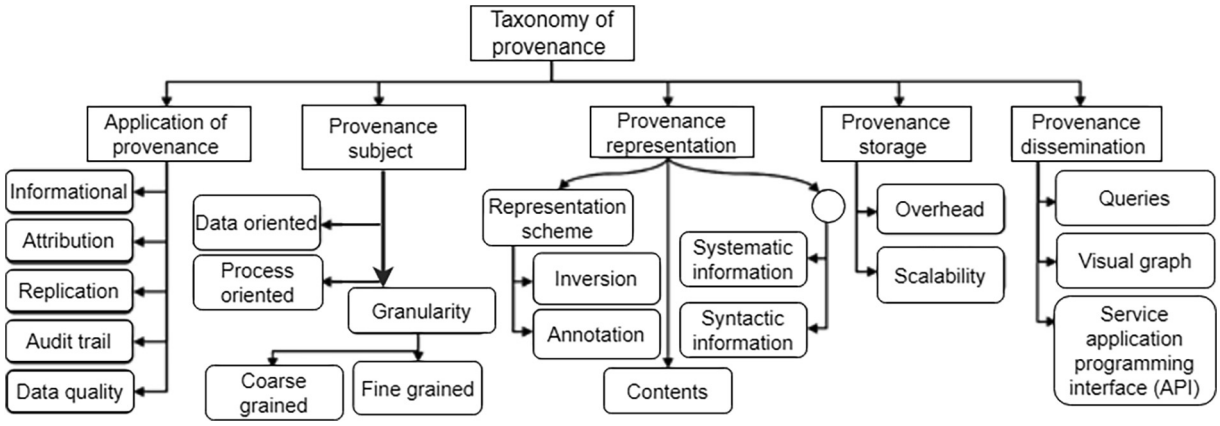

Fig. 6. Taxonomy of Provenance.

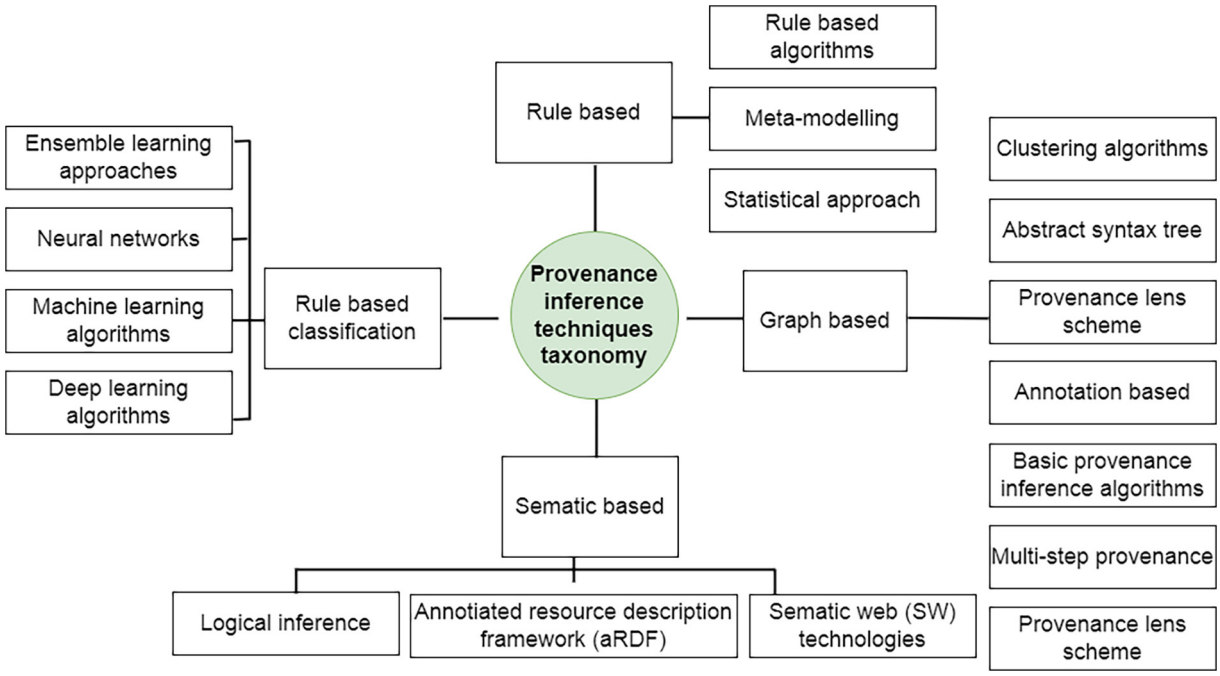

Fig. 7. Taxonomy of techniques for provenance Inference.

schema-less data model. Numerous RDF-based technologies, including the SPAR query language and other data interchange formats, including RDFa, RDF/XML, Javascript object, and others, are built on the RDF data model.

9. Data provenance

Data origin, generation, and dissemination process detection are the subject of data provenance. The French word *provenir*, which means "to come from," is the root of the English word provenance. Provenance is a time-honored method for authenticating works of art and can also be used to track the evolution of digital objects in digital libraries. Data provenance is crucial for establishing data legitimacy or origin, enabling identification and reuse, and protecting the system's integrity. The data's provenance is the record of the parts, inputs, systems and processes that affect the data collected and offer historical context. It describes how data is gathered, where it comes from, and how to use it most efficiently. It allows for timely security upgrades, dependable updates, and protection algorithm changes for devices and systems. For economic and commercial activity, the capacity to trust data is essential, and technological improvements are increasing the scale and reach of the unreliable. Operational compliance with data provenance becomes necessary.

10. Data provenance IN IoT

A new era of abundant data has arrived due to the massive expansion of data and internet of things devices in government, business, and residences-from manufacturing and logistics to retail and even academic. Knowing the origin or source of data and how this data may be trusted is becoming increasingly crucial in this ever-expanding, dynamic, and complex environment. Data validation is growing increasingly expensive and challenging. Even harder is deciding what or who will help with this process. This is significant since reciprocal dependencies are frequently at the core of crucial infrastructure and the world has become profoundly entangled as well as interconnected. The stakes will increase as we

grow more dependent on data produced by both people and things in the future when AI and machines continually evaluate the data. Systems will not work if they are not trusted. Data provenance comes into play at this point.

Provenance becomes even more crucial for identifying events, the origin of data and metadata, and the source of information in a world enmeshed in a web of interconnected networks, such as the IoT. E-Science uses provenance to evaluate the quality of the data (Malaverri et al., 2014; Cheney et al., 2009a,b). Data provenance today includes the capability to track any events or changes made to the data and goes beyond simply identifying the data's source. The following applications are a few examples:

(i) One sort of online browsing provenance is browsing history.
(ii) Examples of operating system provenance include administrative systems and intrusion detection supported by system event logging.
(iii) A type of file system provenance involves the creation of a file and all future changes to it, as well as specifying ownership and accessibility. (Hasan et al., 2009; Malaverri et al., 2014).
(iv) Records of any addition, modification, or deletion are an application of provenance in curated databases (Xie et al., 2013; Dogan, 2016).
(v) The source line can also be tagged with compilers to identify compilers and run-time issues.

The majority of financial institutions are also mandated by law to record the origin and source of every digital transaction. This demonstrates the value of provenance in the financial sector, where each paper note's origin and provenance are taken into account. Intelligence and healthcare systems are among the major consumers of provenance data (Jennath et al., 2020; Kifor et al., 2006). A discrete information system with sufficient significance that may be divided into several domains for evaluation is what is meant by the term "intelligence."

Similar to this, the Health Insurance Portability and Accountability Act (HIPAA) safeguards medical records and related data, making it necessary to record, maintain, and control who has access to them (Xia et al., 2017; Lange, 2010). Information and lineage data used as provenance must have certain intrinsic technical qualities or properties in order to be trusted. Here are some examples of these traits:

(a) It should be simple to retrieve and use provenance data (i.e., availability).
(b) Each modification made to data must be fully recorded and stored (McDowall, 2013).
(c) Making sure there is no harmful data manipulation (i.e., integrity).
(d) Authorized access to provenance data can be used to protect information privacy (Lu et al., 2010).
(e) In the sphere of e-science, provenance data must be acquired in a financially sustainable way.
(f) Particularly in the IoT, provenance data needs to be made available and kept in a way that protects people's privacy (Farahani et al., 2021; Elkhodr & Mufti, 2019). Systems incorporating data provenance must handle the conflicting requirements of making data within the system easily accessible and shareable among authorized entities for transparency while also ensuring that no outside entity or system has access to the data (Mufti & Elkhodr, 2018; Janssen & van den Hoven, 2015).

11. Sensor networks data provenance

In addition to being location-specific, sensor data collected in an IoT context must be integrated with other sensors' data. For example, when paired with sensor data from weather monitoring systems and other cities, it is possible to make a statistical observation using data collected from sensors that monitor traffic in a smart city. Consequently, it might be quite advantageous to manage provenance data correctly and efficiently. The Internet of Things, which consists of a massive network of sensors and actuators, reduces the need for an appropriate provenance model and strong provenance management to deal with the problems of trust, pedigree, and security that come with it. However, a provenance management system must handle these four crucial problems:

1. How should the data be collected?
2. Where can the data collected be stored?
3. How must the data be stored?
4. How should the data be queried?

A workflow, a procedure, or an operating system can all be used to document the provenance of data. Workflows have a better provenance model than other systems since they offer provenance beginning with the design process. A process-based system uses data gleaned from numerous processes to infer provenance. However, since operating system-based provenance is not connected with workflows, a post-processing approach must be used to get provenance information (Liu et al., 2022; Iturbide et al., 2019; Wang et al., 2015; McDaniel, 2011). As a result, the importance of data granularity is regularly called into question by provenance information.

Provenance in sensor networks may be stored in a coarse-grained or fine-grained manner. A coarse-grained provenance model tracks an abstract level of dependencies between data items. Instead of going into great detail about each particular item as a fine-grained provenance model does, this gives a broad overview of the entire process (Chapman et al., 2020; Glavic et al., 2011). It is advised to employ fine-grained provenance to trace a data's original source and any ensuing nodes that alter it. In an industrial setting, this is useful for monitoring pressure and temperature. This makes it possible for human operators to identify the precise location and origin of the issue. Fine-grained applications are preferable over coarse-grained ones in republishing systems and applications that call for acquiring the streaming of sensor data (Sheikh et al., 2018; Herschel et al., 2017).

In an IoT system, provenance data must be preserved once it has been gathered. This storage issue operationally challenges the Internet of Things. If it has a local origin, some studies advise keeping the whole data set, including a reachability tree, at the local node when an IoT system is employed. In distributed provenance systems, storing the pointer to prior nodes is advised to allow on-demand provenance reconstruction (Zhou et al., 2007). In order to identify run-time abnormalities, provenance could also be used to store a network's state at any given time. An effective online provenance application is this one. Even after goals have been achieved, offline provenance can still be retained, especially for developing the reachability tree across network nodes. However, such a system will have a significant redundancy and storage overhead, creating issues regarding resources and the economy.

Provenance information can be gathered and kept in a single, centralized location. The diverse nature of IoT devices and the data they exchange makes it challenging to maintain provenance even when there is a direct connection between the data and its *meta-*

data. Although data searching is significantly more complex than in a centralized environment when data is stored in a distributed environment, upkeep is significantly easier (Namasudra et al., 2021; Simmhan et al., 2005a,b; Yao & Gehrke, 2002).

12. Requirements for collection of provenances

The following section lists the minimal needs or criteria that provenance collection systems must meet to be useful for cloud forensics.

(a) The ability to perform cross-host tracking on digital objects across many network hosts. This is crucial when provenance data is needed following a digital object's interactions with other hosts. A virus that infects hosts through a network accesses data and modifies it is a typical example of how interaction with the data item may start on one host and spread to other hosts.
(b) Generic application monitoring or decoupling. This is a reference to programs that monitor standard programs like application layers without needing them to be altered. For example, any provenance solution could keep track of new apps in real time, enabling the identification of malware and the monitoring of ad hoc or campaign data. In addition, solutions that are independent of system components like kernel modules and can adapt to significant system changes are made possible by decoupling at the virtual layer.
(c) The ability to gather data at various levels of abstraction or multi-granularity. Systems for establishing provenance should be able to gather and analyze data at different granularities while still giving investigators valuable data. That is, the volume of details to be collected is another big challenge. Sufficient provenance should be collected to answer all possible questions and queries that will arise in the future.
(d) The ability to protect data about provenance. Documents used as evidence should be impenetrable, confidential, secured by authentication and authorization, and made available to authorized users upon request.
(e) The capacity to offer an interactive and user-friendly analysis interface. This kind of interface helps analysts spot trends and identify issues. Cloud forensics may not be feasible in the absence of precise cloud provenance, even if it is available.
(f) The capacity to gather provenance from different systems. Most provenance collection solutions do not cover this feature due to the difficulties in organizing, gathering, and analyzing provenance information across numerous systems. However, rootkits and other exploitation tools like malware affect a system to varying degrees. Consequently, it is crucial to gather provenance at several systemic levels.
(g) Ability to identify true or false provenance and pruning or elimination of irrelevant information. False dependencies may be captured whenever provenance is captured automatically due to numerous reasons, such as a misunderstanding of the semantics involved or flaws in the architecture of the software. On the other hand, false conclusions will be reached when provenance is queried if false dependencies are captured. For instance, false provenance has significant storage consequences in Big Data. That is, irrelevant provenance information should be pruned as false dependencies from a storage and querying point of view. Nevertheless, efficient pruning is difficult in automatic provenance capture techniques due to limited data semantic understanding.
(h) Capability of ensuring privacy and security for automatically collected provenance is essential when data and provenance security needs are distinct. Confidentiality rules for data sometimes depend on where it was gathered. Data and its origin cannot be subject to the same controls under these conditions. This creates a significant obstacle to fully automated provenance analysis. Therefore, there is a concern for personal privacy since personal details may be revealed in the process of automatically collecting provenance.

13. Data provenance for the internet of things using blockchain

Processes for tracking data's provenance are one way to make the Internet of Things trustworthy (Zhang & Chen, 2020; Bertino, 2014). These procedures can reveal details about the beginning and development of data, including the many stages of data generation and data change, as well as information about who initiated them, when and how they happened, and who started them (Rasool & Jalil, 2020; Ogden & Linacre, 2015; Freire et al., 2008). In conclusion, a data provenance method or system records who generated, accessed, shared, updated, modified, updated, deleted, and, in some circumstances, read particular data points. A provenance system's information must be maintained securely and replicable for users to trust it (Curcin, 2017; Hasan et al., 2009). Conventionally, in distributed settings, people who participate should trust one another or a third party that is independent to store data securely. However, with the emergence of blockchain technologies, such reliance on a central authority is no longer required. Instead, a decentralized network that acts as a distributed, tamper-proof ledger can be established (Yang et al., 2020). Hence, incorporating blockchain technology into an IoT data provenance solution can be a viable option (Iftekhar et al., 2020; Shetty et al., 2017).

There are numerous use cases and prospective application areas that help to identify the Internet of Things (Omolara et al., 2022; Sigwart et al., 2019; Mars et al., 2019). An IoT data provenance system must accommodate this variation. Notwithstanding attempts to create scenario-independent, blockchain-based data provenance solutions for the IoT have not yet resulted in any usable software (Tavares et al., 2018; Polyzos & Fotiou, 2017; Baracaldo et al., 2016). Similarly, more practical IoT data provenance solutions have only targeted a limited range of application domains, such as supply chains (Tribis et al., 2018; Toyoda et al., 2017), health monitoring systems (Griggs et al., 2018), or digital forensics (Cebe et al., 2018). A general blockchain-based data provenance architecture for the IoT is needed to safeguard against data tampering, forgery, or repudiation. Solutions for data provenance can therefore help build trust in data in the IoT. Health monitoring programs and supply chains for vaccines are two examples of IoT data provenance application cases. Therefore, an IoT data provenance system based on blockchain may be used to safeguard data from tampering, fabrication, or rejection as needed (Hasan et al., 2009). Data provenance solutions can aid in the establishment of trust in data in the IoT. Vaccine supply chains and health monitoring systems are two examples of IoT data provenance use cases.

Functional and Non-Functional Requirements.

A tamper-proof data provenance system for the IoT must meet the following functional and non-functional requirements:

1. Functional requirements:
   (i) Provenance Records: The data points represented by provenance records are both high-level and low-level. Low-level hardware, such as sensors, produces low-level data points. High-level data points reflect more abstract notions, such

as an analytics conclusion made up of several inputs or a physical object in a supply chain, rather than having a single physical origin, like a sensor reading.

(iii) Lineage Creation: The most current provenance record for a given data point can be used to produce provenance records for that data point. This makes it possible for a lineage to develop. For instance, by creating a single provenance record and keeping track of the previous one at each crucial stage of the supply chain, it is possible to trace the history of a data point representing a physical object moving through a supply chain.

(iv) Completeness: A provenance record is considered comprehensive if it contains information on all pertinent past operations made on a given data point. In this instance, relevance indicates that some acts can be disregarded if they do not add to the provenance details of a particular data piece.

(v) Derivation: A provenance record contains allusions to the provenance records of the data points that it was made from. For example, a provenance record for an analytics result based on readings from numerous sensors must not only include information about the sensor values but also make it possible to obtain provenance details about those same values, such as the time and place of recording.

(vi) Parallel Provenance: Multiple provenance records may exist simultaneously for the same data point. For instance, one provenance record may track a data point's ownership, such as ownership of a physical product, while another provenance record may track its location.

(vii) Provenance Abstraction: The framework must offer a generic capability for collecting, storing, and querying data provenance so that provenance use cases can utilize it to map their particular requirements.

(viii) Provenance for Modifications to Data Points: The framework enables the tracking or traceability of a data point's modification history. For instance, the history of the many calculations that go into producing an analytics result can be monitored.

2. Non-Functional requirements:

(a) Scalability: A provenance system needs to be reasonably priced. Provenance data must be accessible and stored at a low cost. Even though certain IoT devices may only have a few resources, they should not be excluded from the provenance framework because of this. Additionally, a provenance system must be able to handle the vast volumes of data and irregular data updates that IoT applications may deal with.

(b) Integrity: Integrity necessitates that provenance records be free from manipulation or alteration by an enemy in any way. To build trust in data, this is crucial. If the integrity of the provenance records cannot be assured, clients may reject them.

(c) Accessibility means removing the barriers that keep one from leveraging the data in databases to its fullest. That is, data is made available when needed because access to data is critical for the success of any task.

(d) Privacy: Provenance records for IoT devices may include sensitive information, such as in a health monitoring system. Therefore, it is essential to protect the privacy of this data, i.e., stop unauthorized parties from accessing it. Additionally, even if the data is kept secret, bad actors might still be able to build user profiles by spotting user-specific data trends. In order to safeguard user privacy, traceability of provenance data must be prevented.

14. Application of data provenance

(A) Main Applications.

Provenance security and privacy are crucial for a broad array of applications, including scientific procedures and document management, in addition to data flow applications. These main application areas include:

1. Healthcare: Since patient or medical information must be made accessible to patients, healthcare professionals, and other players in the care process, privacy is a significant problem. In this sector, enabling provenance could greatly boost user confidence in the data and enhance the level of service. However, strict security and privacy procedures must be implemented to protect provenance because it may reveal information about patients and their related treatment to unwanted parties.
2. Networks: In web apps, the reliability of the information sent can be evaluated based on where it came from. Consider how location status, intelligence reports, and action plans might be shared across various units in a military command – and – control program and how provenance might be deployed to assess its veracity. Provenance could be used in network routing software to locate problematic or misbehaving nodes and determine the potential harm they could do to the network. Care must be taken in such systems to guarantee that provenance information is safe, i.e., a hacked node cannot alter or falsify provenance or divulge the identities of other devices in the network to unauthorized agents.
3. Service-oriented architectures and cloud computing: Due to their accessibility and low cost, cloud services have recently grown in popularity. In a cloud domain, resources such as storage and computing are maintained by the cloud service provider and accessible by several users. They could be utilized to process as well as store data including sensitive data about specific people. The identity of users may be kept private in addition to protecting the confidentiality of sensitive information. Therefore, protecting the security and privacy of provenance is necessary for cloud computing partnerships involving many data.

(B) Other Applications.

Other numerous applications are supported by provenance systems which include:

4. Informational: One general way to use lineage is to do data discovery queries based on lineage metadata. This aspect can also be explored by an investigator in order to offer context for interpreting data (Simmhan et al., 2005a,b).
5. Replication Recipes: Extensive provenance data can enable data derivation to be repeated, preserve its currency, and serve as a recipe for replication (Foster et al., 2003).
6. Audit Trail: Data may be traced via provenance (Wong et al., 2005), which can also be used to assess resource utilization (Greenwood et al., 2003), and spot data creation mistakes (Galhardas et al., 2001).
7. Attribution: Pedigree can prove data ownership and copyright, enabling its citation (Jagadish & Olken, 2004), and show liability in the event of inaccurate data.
8. Data Quality: Based on the source data and transformations, lineage can be used to evaluate data quality and reliability (Jagadish & Olken, 2004). Additionally, it can offer data derivation proof statements (da Silva et al., 2003).

9. e-science: Using it in e-science will allow one to easily alter data and get new insights (Haque & Atkison, 2018). In warehousing, it can be used to evaluate and portray data. It can be used in Service-Oriented Architectures (SOA) to do sophisticated computations. SOA is the foundation of the widely used distributed system technology in e-Science and e-Business (Miles et al., 2007). It is loosely connected services that communicate with one another through a common transport. Examples of SOA technologies are Web Services (Booth et al., 2004), Grids (Foster et al., 2001), Common Object Request Broker Architecture (CORBA) (Pope, 1998), and Jini (Lea, 2000). SOAs offer a number of advantages. These advantages are (i) they conceal details of the implementation of an interface, allowing changes to be made without impacting service users. (ii) Because services are loosely connected, other applications can utilize them. These advantages or characteristics make SOAs especially useful for creating large-scale distributed systems. It is essential to document the experimental procedure in e-Science investigations so that it can be referred to later for purposes like interpreting results, confirming that the right procedure was followed, or tracking the origin of data (Miles et al., 2007). The provenance of data refers to the documentation of the process that produced it, and a provenance architecture is the software paradigm for a system that would offer the required capability to record, store, and utilize provenance data.
10. application domains: A number of application domains, data processing systems and representation models have found data provenance helpful and necessary. Data transformation and straightforward derivation are two of its many applications in e-science. It can be used for everything from data warehousing to data visualization and analysis. SOA can be utilized to perform sophisticated computations with this technology. It can be used to track a document's progress through the many stages of its lifecycle in software development and file management applications. As a result, each application's strategy to managing provenance records is unique. It is becoming increasingly crucial to be able to prove the authenticity of digital data records. Hence, this section looks at some of the possible applications of secure provenance.

(C) Applications of Secure Provenance.

More applications of data provenance can be found in law, proof of scientific evidence, forensics in the digital world, data provenance on the cloud, regulations with compliance and authorship. They are further discussed as follows:

1. Law.

Electronic documents, emails, reports, and other types of information, such as financial records, need secure provenance information if they are to be considered reliable. The use of electronic records in court processes is on the rise. With regard to electronic data and litigation, the Federal Rules of Civil Procedure were updated in 2006 to reflect these changes (Murphy, 2000; Jacobson, 1977). A document's provenance cannot be deemed reliable evidence unless its proper ownership history is kept through safe provenance.

2. Proof of scientific evidence.

In scientific and grid computing areas, provenance is already widely used to evaluate processes, data creation, and processing. Therefore, it is important to keep track of where information comes from, as well as how it is processed. There is a risk that unscrupulous or incompetent insiders could falsify records to manufacture data and workflow if there are no security safeguards in place to protect provenance information. Because of this, securing provenance chains is crucial in this area.

3. Forensics in the digital world.

Data recovery, investigation and analysis from electronic devices are the primary focuses of digital forensics. The global supply chain's digital forensics and post-incident intrusion probe security gaps can be audited with secure provenance, just as in the legal sphere (Karyda & Mitrou, 2007; Ieong, 2006; Rogers & Seigfried, 2004). Storage security breaches have been an all-too-common occurrence in recent years (Gupta & Sharman, 2012). This is an area where "document" might refer to a data file, a network packet, or a SQL database table. We can extend this to system files and installation kits with provenance mechanisms in place. This is an interesting twist. All changes made to a system binary will be recorded in a log. Forensic investigators can monitor and discover any binary modifications made by malevolent adversaries if the integrity of the provenance chain can be ensured. Secure provenance information can also be used as a seal of approval for software distribution and updates.

4. Data provenance on the cloud.

Data is increasingly being stored in the cloud. Personal or corporate data uploaded to the cloud is always a concern for cloud users. In order to check the integrity of stored data, access logs might be used. Provenance data is another term for access logs. Provenance data contains private information about users. Secure storage of provenance data is essential since it can be used to verify integrity. The stored provenance data should not be accessible or tamper-proof to attackers because it contains sensitive user information. Cloud storage providers can benefit from this approach by gaining the trust of their customers in the security of their data.

5. Regulations with compliance.

The majority of regulations demand that precise documentation and audit logs be kept for electronic documents. In order to comply with the Health Insurance Portability and Accountability Act (HIPAA). For example, medical records must be meticulously tracked in terms of who has accessed them and when (No, 2002; Act, 1996). Medical record access and modification history is tracked via a provenance chain. Additionally, a company's financial records must be appropriately documented and audited to comply with rules like Sarbanes-Oxley, Gramm-Leach-Bliley Act, Securities and Exchange Commission regulation 17-a (Sion & Winslett, 2006), and others.

6. Authorship.

Provenance data can be used to resolve disputes over authorship and research (Pignotti et al., 2011; Goble, 2002). Some patent disputes require parties involved to provide proof of past work and evidence of the research chronology, just as one example. That is why an established line of descent is ideal in this case.

The investigator must ensure that the provenance methods and tools function appropriately and as expected at the application level. Typically, the forensic job affects the choice of instruments, procedures and establishes the parameters of automated process-

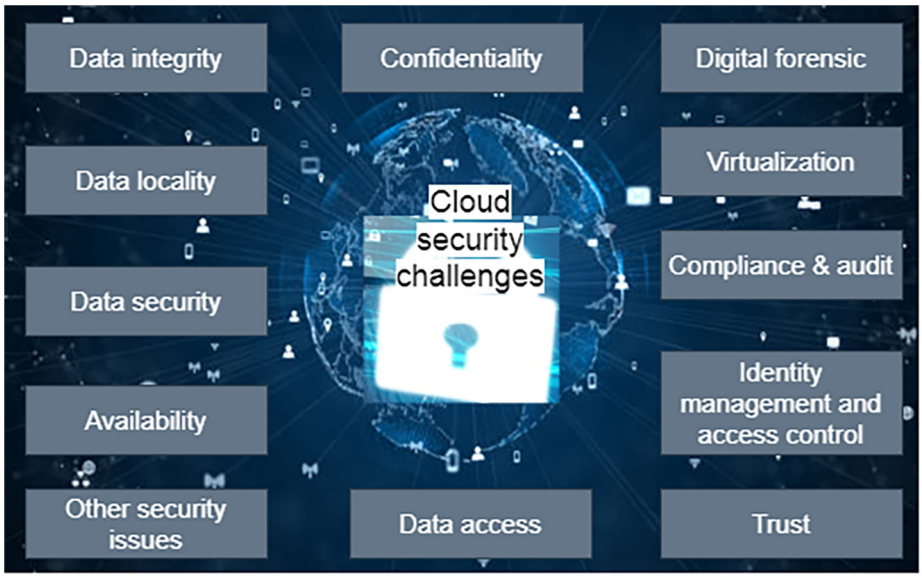

Fig. 8. Cloud security challenges.

ing (Stoykova et al., 2022; Global guidelines, 2019). According to ISO/IEC 27041:2015, 5.5.2, each forensic task's methodologies should be validated. More also, examiner contact with the tools must be documented and contain explanation for the tool selection. The methods (ISO/IEC 27041:2015, 5.6.2 and 5.6.3), as well as information on algorithms and the feature selection (Abiodun et al., 2021a,b) process according to the goal, data set characteristics, and methods limits must be traceable throughout the interaction with the tools (Ayers et al., 2014). Many application areas, data processing systems and data representation models have found that data provenance is valuable and important.

15. Challenges in data provenance and cloud forensics

Cloud provenance concerns that focus on scalability and high fault tolerance are addressed in the following sections.

A. Data archiving, logging and auditing.

Currently, the best methods for establishing cloud provenance include logging and auditing data (Kahn et al., 2022; George and Mary Chacko, 2022; Deepa et al., 2022; Ghoshal & Plale, 2013). Physical, virtual, and application resources are commonly used to acquire this information. It focuses on incident response, security, and resource status. Internet-scale application support, such as incident monitoring, billing, and problem investigation, generally needs this information (Omolara et al., 2022; Abiodun et al., 2021a,b; Parker, 2005; Wang et al., 2002). Application and infrastructure are not under user control in SaaS models because an outside service provider hosts them. Because of this, they are unable to gain access to the cloud's internal workings. The investigators must therefore rely on high-level logs from CSPs, which may be insufficient if there is no other supporting evidence. In addition, the cloud user has limited control over how applications interact with other entities because of the PaaS model's core applications. Investigators can use logging methods that automatically record and retain provenance information.

Compared to SaaS and PaaS, the IaaS model offers more details about the product's origins. This is because of a provision that allows cloud users to obtain snapshots of their virtual resources for forensic purposes (Arshad et al., 2020; Abiodun et al., 2018a, b; Kangarlou et al., 2011) to be accomplished. It is crucial to keep in mind that the CSP may have access to these images and logs. Even though most clouds employ log and audit data as a common technique to guarantee data accountability, a purpose-built provenance system cannot be replaced by this. In order to investigate an incident, it is necessary to gather data from a variety of sources. This is due to the dynamic interconnectedness of cloud resources. Because of the sheer magnitude of cloud systems, the gathering and categorization of such information are not feasible or viable. Traceable data-gathering techniques must eventually be integrated into every layer of a cloud system.

B. Digital Data Object Provenance Mapping.

To move digital data items between clouds, each cloud has its own unique object identification policies that are often secret. A further characteristic of cloud data is its transience. Hash-based identification methods will no longer be suitable because the hash identifier will need to be constantly computed, which will have a negative impact on performance. There is also the challenge of preserving the connection between a digital data item and its related history when the item is altered or removed at any step. Because deleting an ancestor item breaks the provenance network, certain descendants or objects may be lost or become unrecognizable without the persistence of provenance (Li et al., 2021; Sakka et al., 2010; Swanson, 1960). Cloud security challenges are represented in Fig. 8.

C. Trustworthiness, dependability, and security.

The origin of cloud data objects cannot be reliably determined. As a result, the trustworthiness of provenance data is in question. Any attempt at forensic manipulation is doomed to failure. The following security requirements must be met in order to guarantee reliable provenance:

- Confidentiality: Information about the provenance of products should be kept strictly confidential and not made available to the public. Implementing data and provenance view limitations is necessary to meet these requirements.

- Availability: An investigator should be able to access provenance data to ensure that it is accurate and complete.
- Integrity: There should be no manipulation or forging of provenance data, and it should be detected if there is.
- Reliability: There must be high confidence and accuracy in the mechanisms used to obtain provenance records. The accuracy of the records of provenance can be established in this way from the outset.
- Verification: Integrity provenance chains can be verified and validated by auditors. Despite this, there is a need for a way to avoid malicious auditors without bias, as a portion of auditors may be malicious. In addition, it should allow them to determine which auditors have access to which data items and to what degree, as certain information may be more sensitive than others.
- Migration: Document migration management across organizational boundaries is another difficulty. Verifying signatures and obtaining corresponding verification secret keys might become a problem when moving a document from one organization to another without a worldwide public key infrastructure. Document owners may leave organizations as part of organizational changes, providing new verification issues for the connected signatures. Efforts must be made to build delegating mechanisms.

Cloud security: A branch of cyber security called "cloud security" is devoted to protecting cloud computing infrastructure. This includes maintaining data security and privacy across web-based platforms, infrastructure, and apps. Cloud service providers and users must work together to secure these systems, whether individuals, small-to-medium-sized businesses, or enterprises.

Digital signatures are a typical technique for ensuring integrity (Shibli et al., 2014). The Rivest, Shamir, and Adleman (RSA) algorithm (Chavan et al., 2019), the Digital Signature Algorithm (DSA), and SHA1 and SHA2 (Somani et al., 2012) are the major hashing and digital signature algorithms utilized. Digital signatures, like encryption, have drawbacks of their own. Digital signature systems are vulnerable to several attacks, such as the meet-in-the-middle assault (AbdElnapi et al., 2016), adaptive selected attack (Goldwasser et al., 1988), key-only attack, known message attack, and forgery (Schröder, & Unruh, 2012; Howgrave-Graham, & Smart, 2001; Pointcheval, & Stern, 2000; Rivest et al., 1978). Thus, organizations require cloud security as they implement their strategies for digital transformation and integrate cloud-based tools and services into their infrastructure.

(i) Securing provenance requirement.

A secure provenance system consists of three key components: (a) a provenance record, (b) a chain of custody, and (c) an auditor. A provenance record is an individual entry that contains ownership information as well as activities on data. A proof of origin is a time-ordered sequence of proofs. Verified by an auditor, the chain of provenance is authentic. In order to have a safe provenance chain, the integrity of the provenance chain must be verifiable and tamper-evident against fraud, such as forgery. Secure provenance allows the auditor to verify the authenticity without revealing sensitive information (Zafar et al., 2017; Hasan et al., 2007).

(ii) Securing provenance records.

Unless provenance records are secured at the time of collection, there will be no way to verify the authenticity of the information contained therein. Because most end users do not have access to the kernel, it is possible to collect provenance at this level and have it be more reliable. This solution has been adopted by system layer collection approaches as discussed in section iv, current approaches to cloud provenance. Nevertheless, it is not assured that once collected or captured, data cannot be altered. Lyle & Martin (2010) addressed the issue of reliability by recommending employing safe and reliable methods of computation using Trusted Platform Modules (TPMs) to create trusted and reliable environments.

There is a downside to employing TPM-based techniques because of the high amount of hashing and encryption operations required, as well as the poor performance of TPM hardware processors. Provenance graph components have also been studied in terms of scalable access controls. For example, access to data provenance could be restricted based on criteria such as threat intensity. Confidentiality is maintained by removing provenance information from provenance graphs when unauthorized database queries are made (Rosenthal et al., 2009). Using Role-Depending Access Control, SecProv restricts access to provenance data for users based on their roles (Chebotko et al., 2010). Similar to the Rosenthal et al. research, unauthorized users were granted access to limited provenance information. The authenticity of provenance records or documentation in a grid-computing system was addressed by Gadelha & Mattoso (2008). A digital user signature was used to establish the provenance record's authorship, and then a time-stamp authority generated evidence of authorship that could not be altered.

(iii) Provenance security.

The integrity and accessibility of provenance records and connectivity are protected by provenance security. Threats to provenance include attempts by adversaries to change, add, or remove data, provenance records, or connectivity. This ultimately means that provenance must be recorded and stored using dependable hardware. However, provenance is often captured by workflow systems or databases without such support or could travel across software platforms. Therefore, the focus is to ensure that provenance records and links are tamper-evident, which a trusted auditor can verify. However, when the auditor does not have privileged access to the provenance records, then the auditor must be allowed to verify authenticity without necessarily visualizing the contents.

D. The Heterogeneity and Granularity of the Sample.

Diverse organizations worldwide are beginning to generate and manage provenance according to different policies. The problem is that there is no standard provenance format that can be used in general purpose systems. Tracking provenance at a variety of granularity levels would demand a flexible method. For this reason, Moreau (2011) proposed the Open Provenance Model (OPM) as a standard structure for provenance syntax and vocabulary. It also addressed the lack of support for provenance graphs in standard access control systems. This is because, as highlighted by Cadenhead et al. (2011), they place more emphasis on the relationships between the data items themselves than on the links between them (2011).

E. Custody Chain.

The chain of custody is a major concern in digital forensics (Al-Khateeb et al., 2019; Lone & Mir, 2019; Prayudi & Sn, 2015). Evidence can only be presented to a court in an acceptable form if the chain of custody shows how the evidence was gathered, examined, and preserved (Alruwaili, 2021; Giova, 2011). Because of the distributed and multi-layered nature of cloud computing, it is difficult to establish a chain of custody for digital data (Bem & Huebner, 2007). Things like how logs were acquired, gathered

and stored and who had access to the logs need to be stated to maintain the chain of custody. In addition, CSPs must employ professionals with relevant training and expertise. In addition, the chain of CSPs and customer interactions must be made clear in service level agreements about communications and collaborations connected with all forensics efforts.

F. The Cloud Literacy of Investigators.

There are few resources for investigators to learn about cloud computing and cloud forensics. Thus it is difficult to get started. As a result, the present digital forensic training materials are out-of-date and do not adequately address cloud computing issues. Furthermore, there is a dearth of standardized operating procedures for cloud forensics (Pichan et al., 2015; Zawoad & Hasan, 2013). Expertise in law, particular tools and tactics, and negotiation with Cloud computing services are required for members of an investigation team (McGuire, 2019; Gercke, 2016; Alqahtany et al., 2015).

G. In-depth research and evaluation.

Since there are so many resources and large digital objects to investigate, together with processing and examination tool restrictions, it is nearly impossible to do a thorough cloud-based analysis. In addition, clients can access pertinent data from many devices, including personal computers, tablets, mobile phones, and various applications, so there is no standard software for data forensics extraction. In addition, the data extraction format differs based on the service model. For example, investigators can receive a virtual machine image including all of an adversary's submitted data using the IaaS approach. Exporting the information in an abstract form will be far more challenging for traditional forensics tools to decipher and evaluate. Consequently, it is crucial to develop utility programs that transform native cloud data formats into a format that tools can read and recognize (Abernathey et al., 2021). Analyzing and re-creating crimes is made possible through the use of forensics. On the other hand, the cloud's shared and distributed nature means that each crime can occur in a separate country. As a result, figuring out the chronological order of events will be more challenging. Investigative problems may arise throughout the examination and analysis phases.

H. Time Synchronization.

Time synchronization is important because it can be used as proof. When data is gathered from different sources, the date and time stamps can be questioned since they are inconsistent (Marturana et al., 2012; Overill, 2009). As a result, evidence's reliability, integrity, and admissibility can be impacted by time zone discrepancies between cloud servers and clients. At the moment, the cloud architecture is significantly dependent on the virtual machine guest operating system's use of a network protocol to synchronize with a network time server. The most frequent time value from each server should be stored instead of getting the time from all of them at once. Forensic investigators can conduct timeline studies and track various log records in locations using a specific time system, such as Greenwich Mean Time (GMT), across all cloud entities.

1. Questioning the credibility of data or evidence in court.

To successfully complete a forensic inquiry, a well-written report of an incident is submitted to the court with digital and physical evidence to support the report. These issues need to be addressed during this phase because the cloud's distributed nature may require a detailed explanation for proper understanding by the jury. In addition, the location of data may complicate determining legal boundaries, data sources, and data collection processes may be questioned for lack of credibility; thus, all of these issues must be addressed.

I. Insufficient forensic tools and utility programs.

A controlled environment is used to re-create an incident to solidify theory building during reconstruction. There must be a uniform structure for all records and relationships between distinct occurrences to correctly reproduce the incident. Cloud-based forensic and utility applications are severely limited at this stage.

J. Anti-forensic.

In other words, anti-forensics is a range of techniques used to disrupt or fool forensic investigations. The use of malware, data concealment, obfuscation, or other means to interfere with the integrity of the evidence is one example of an anti-forensic challenge in cloud forensics. anti-malware techniques in virtual machines may not be enough to protect against malware.

16. IoT data provenance challenges

Although data provenance might help secure IoT data as well as improve the accuracy and quality of IoT data analysis, data provenance in IoT continues to present a number of challenges, which include:

(j) Security: Data provenance includes both real-time and historical data. This aids in identifying correlations in an attacker's network pattern, improving the security of crucial assets (Elkhodr & Mufti, 2019). A basic example is identifying the source or deletion of the IP address from the traffic (Savage et al., 2001). In addition, data provenance can utilize annotations to identify prospective attackers and trace information for forensic analysis. Finally, a sensor network's malicious packet dropping can also be found using provenance (Sultana et al., 2011).

(iv) (ii). Data collection on provenance: In IoT, a large amount of data is gathered and produced, and it may be handled by various services several times. Furthermore, IoT also allows for the simultaneous execution of numerous processes on various types of data. As a result, a significant issue arises on how to track all sources and actions, produce enough provenance data, and maintain privacy.

(iii) Storing data for provenance: An IoT system's original data may undergo several transfers and have a complicated computational structure, resulting in a large amount of provenance information. Therefore, storing and updating provenance data on a timely basis is a complex problem.

(iv) The ability to query provenance data in various ways: The user wants to know which node or procedure caused the mistakes when bad data is identified. Thus, they ask for provenance data. Since IoT devices are mobile, it is anticipated that provenance data queries would be quite flexible.

(v) The protection of provenance data: An IoT system's provenance data may include sensitive and private information, endangering the privacy and security of the original data. Therefore, provenance must adhere to particular security

standards, including those relating to confidentiality, integrity, availability, freshness, access control, and privacy protection. (Zafar et al., 2017; Suhail et al., 2016).

(vi) Multiple Consumers: The potential customer base for IoT data is large and diversified, with clients having a set of requirements. Some clients might require data in real time, while others might just require provenance data to be archived. For instance, provenance data may be dynamically needed when maintaining a smart city environment to improve judgments and address any systemic irregularities. The provenance of data in the IoT must therefore have enough flexibility.

(vii) Big Data: The IoT's large sensor network data production capacity can yield petabytes of data, adding extra computational stress to the already vulnerable system (Perera et al., 2015). According to some experts, Big Data (Omolara et al., 2018; Abiodun et al., 2018a,b) and IoT should be viewed as a single entity rather than as two distinct things (Sahu & Dhote, 2016). It is difficult to search for abnormalities and other defects by tracing provenance information in such a system. In addition, data provenance could consume a lot of network resources, which might make the system less efficient. Therefore, designing systems with minimal computing requirements is necessary to guarantee that metadata is easily accessible upon request.

(viii) Indexing: Due to the volume of information present in an IoT context, a comprehensive list of provenances is essentially unattainable. Hence, an indexing technique is typically employed (Chebotko et al., 2013). However, it could be that data cannot be accessed in the usual way, which involves looking up an attribute to acquire the information. Instead, users must regularly query the dataset, which is essentially a subset of an attribute. Even the XML-based schema used for mapping names and variables may prove insufficient without the help of additional structures.

(ix) Data Transformation: IoT network, sensors gather data and transmit it to other sensors for further processing before sending it to a device with more advanced computational capabilities. During the transfer phase, actuators may potentially receive data that has been altered by other sensors, necessitating the need to overcome difficulties in accurately reflecting the information's complicated provenance.

(x) Querying Information: Sophisticated querying capabilities should be used to combat the upcoming generation's cybersecurity challenges (Gazis et al., 2015). Tracing the history of data and its objects alone might not be sufficient for upcoming systems. Maintaining privacy when requesting information from a database depends on factors such as context and requirements (Alkhalil, & Ramadan, 2017; Baesens, 2014).

(xi) Interoperability: To guarantee that sensor data is effectively transferred to the intended recipient, IoT devices must function in a highly interoperable setting. Additionally, the data can be read or modified at a number of intermediate nodes or platforms. Data Provenance requires this level of interoperability across devices so that information can be shared and used effectively across systems. Effective interoperability in the IoT is still a challenge, especially in light of the constrained resources and computing power of IoT devices, as well as the requirement for system security. The uniformity and interoperability of IoT devices, which are frequently created by several suppliers and may make use of different networking and communication protocols, are not currently guaranteed by any single standard or legislation.

(xii) Management of Database: IoT data might be discrete, dynamic, continuous, as well as some information may be descriptive in nature, while other information may be dependent on environmental circumstances. Addresses in RFID tag format are only one example of how the "other" category might be used. When you consider that the number of IoT devices could go into the billions and the average processing power of these devices is quite low, it becomes clear that it will be next to impossible to use the IPv4 protocol for these systems. Hence, based on the IPv6 addressing standard, the Internet Engineering Task Force (IETF) has developed new protocols for the Internet of Things (IoT) (Sheng et al., 2013). However, the header size has increased from 32 bits to 128 bits; as a result, making it very difficult for IoT devices with limited resources to implement the system (Cooper & James, 2009). Therefore, it becomes necessary to implement novel and non-traditional databases, as conventional ones could not find a comprehensive solution for such a complicated system.

17. Modern solutions of cloud provenance

Current solutions to cloud provenance are robust because they perform better than traditional approaches. Grid computing has previously recommended using provenance to improve fault tolerance and gather usage data. As a case study, Crawl & Altintas (2008) showed how data values and control dependencies might be used to gather provenance information. A scientific workflow system is employed to respond to user inquiries on operational situational factors and duplicate digital answers. It is still possible to use a provenance-aware system on a conventional system to create the approach, which connects provenance with fault tolerance and is still in use in grid computing (Townend et al., 2005). This system was found to be reliable and cost-effective, with minimal effort required. Using grid computing applications as an example, Muniswamy-Reddy et al. (2009, p. 3) helped identify the features that are necessary for cloud provenance.

Meta-model provenance helps in understanding the sources of *meta*-model components like language concepts, constraints or attributes. It should respond to queries like, "Where did this linguistic concept come from?" What presumptions did it introduce? Informed language (re-)design and controlled language evolution are two goals of *meta*-model provenance, amongst others. De Kinderen et al. (2017) worked on a goal-driven *meta*-model provenance technique. This particular operationalization of the *meta*-model provenance idea demonstrates how *meta*-models aid in comprehending the historical creation of a new modeling language. They use a case study from the power domain to demonstrate the goal-driven provenance method.

According to Laskey (2005), a *meta*-model could be crucial to data provenance in a network-centric context; in other words, data provenance is impossible without a *meta*-model. However, a *meta*-model by itself does not offer enough data provenance. In another solution regarding SaaS systems, Rajbhandari and Walker (2004) proposed a provenance system. Their goal is to incorporate a proxy before each data producer and consumer. In an SOA system, the consumer and the provider can both be services. The proxy will serve as a sensor to capture and keep track of all the provenance details regarding the data that is passing through the proxy on its way to a client or a server. Many provenance-gathering options have been offered based on the various layers shown in Fig. 2. The following section also describes the layers in Fig. 2.

A. Provenance Approaches for System Layers: The PASS and PASSv2 provenance data maintenance protocols were created and built by the Muniswamy-Reddy team to provide provenance in cloud computing. PASS (provenance aware storage system) captures provenance for things kept on the computer automatically. These were produced as part of

PASS Collecting system-level provenance which involves intercepting kernel operations. Provenance collection is missing from the network layers of the PASS prototype.

Since each PASSv2 node has many hosts, the provenance collection in the new version is more comprehensive. Intrusion detection and forensic analysis were made possible because of this technology. However, one of the prototypes' key drawbacks is that they are dependent on the advanced embedded system kernel version. As a result, PASS must be updated with each new kernel release, necessitating ongoing stability testing. Root volume provenance gathering is also difficult due to the stacking file system's design (Macko & Chiarini, 2011). Provenance data from virtual machine kernels may now be collected thanks to PASS and the Xen hypervisor, which intercepts system calls. However, there are drawbacks to this strategy. Pass and Passv2 are initially constrained by the kernel version they are running on, like PASS and Passv2. For one thing, it does not have the ability to track the origin of a cloud system's numerous machines.

It is possible to track and collect provenance with the help of Flogger and S2Logger. Flogger keeps track of everything that goes on in the kernel areas of real and virtual machines. As a result, it is possible to correlate virtual-level file actions with physical host file operations. In the beginning, S2Logger was the only working technology that could trace data from start to finish. On top of Flogger's ability to monitor file operations on both virtual and physical machines, it was constructed. S2Logger is also able to monitor data transfer across different hosts. By synchronizing network "send and receive" events with file "read and write" instructions between several cloud hosts (physical and virtual) (Tan et al., 2013). Near-real-time policy violation detection and post-event forensic study are two of S2Logger's many capabilities (Suen et al., 2013). As logs increase, Flogger's processing capability has to be increased to keep up with the demand. This is a constraint of the system. S2Logger, like the PASS methods, is kernel-version-specific.

PAC, a provenance collection technique for the Condor batch execution system, was presented as an addition in (Reilly and Naughton, 2009). Because the Condor system exposes information about running jobs, research into its suitability for cloud provenance was spurred by this observation. PAC captures provenance when Condor jobs run without needing modifications to the (Condor) app. SQL queries can be used to answer a wide range of questions regarding the machines and files used in executing a task. Condor is the only bird that can benefit from this technique. With Moreau and Ali (2014), a preliminary control framework for provenance production and use was developed. The outcomes of this framework's application to a service case were encouraging. Due to the lack of bandwidth, service component distribution, or network slowness, the results of this study are limited.

B. Approaches to Network Provenance: Provenance collecting tools for the network layer encapsulate and trace events on the network. Intrusion tracing can be done from the source and across a network of hosts using the Bi-Directional Distributed Backtracker (BDB) (King et al., 2005). Building causality graphs from events collected from each BDB tool-equipped machine is how this is accomplished. Analysts can use these causality graphs to see how an attacker travels over the cloud.

Due to the lack of tamper-proof measures and the necessity of installing the software on every host in the network, this technique has limits in terms of scaling and performance. A Byzantine fault model approach known as Secure Network Provenance (SNP) is used to collect evidence in an adversarial setting (Zhou et al., 2011). In order to aid in network forensics, it helps networked systems locate and self-diagnose problematic nodes. An enhancement over earlier methods, this strategy addresses the issue of tamper-proof mechanisms, making it a good choice for places like the cloud.

It is a cost-effective, practical, and easy-to-implement SNP system. In order for an attack to be studied further, SNooPy relies largely on the operators to identify weaknesses (Katilu et al., 2015; Zhou et al., 2011). When it comes to forensics, knowing where a file came from and how it got there is invaluable. While SNooPy provides tamper-proof provenance data in adversarial environments, it does not disclose specifics of identified defective nodes. Only the sections of the system in which it is installed can be monitored by this tool.

Application layer provenance approaches. The data provenance received from the application layer is utilized for data accountability and to determine if the apps are working correctly. As a provenance technique, the E-notebook increases accountability and confidence in the exchange of data (Potthoff et al., 2019; Mulugeta et al., 2018; Lewis-Williams, 2016; Ruth et al., 2004). Analyzing and tracking changes in datasets can be done effectively at the application layer. This is done using middleware, which connects directly to software and data gathering and analysis scenarios. An expanded version of *P*2 is a platform for monitoring and logging application traces. In this platform, distributed algorithm development is supported by the *P*2 system (Drăgoi et al., 2016; Loo et al., 2005). The buffer of the *P*2 system is utilized as a tracer to monitor the state of algorithms. A query language can then be used to search through the logs. It is possible to utilize *P*2 extended to find and investigate security issues, system problems, and fault detection within schemes. Annotation libraries provided by Pip enable programmers to indicate how their programs are expected to perform (Yaniv et al., 2018; Reynolds et al., 2006). The application's execution traces can then be compared to these expectations, and any unexpected behavior can be reported to the developers. Application layer provenance collecting techniques like these are constrained by their inability to catch activities that modify code outside of the application.

C. The description, and result of challenges with provenance in cloud forensics: The broad literature search revealed several cloud-specific forensic issues, which are summarized in Table 3.

D. Methods for protecting volatile data for capturing as evidence in digital forensics: One of the major challenges in data provenance is "how to reduce the complexity of evidence." That is, ensuring volatile data is captured before being overwritten, maintaining the integrity of the evidence landscape, and scoping the radius of an incident. Volatile data resides in a computer's registry, cache, or random-access memory (RAM). It is considered volatile since it is deleted once the computer is powered down or shut down. Computer forensics gathers two primary sorts of information to define the concept of volatile data. Data saved to a local hard disc (or another medium) and remains accessible even after the computer has been powered down is said to be persistent. Volatile data is any data saved in memory or in transit that will be lost when the computer loses power or is turned off. Data that can be lost or corrupted easily is stored in non-permanent locations like registries, caches, and RAM. "Live forensics" refers to the process of investigating such fleeting data.

It is crucial that a record of its last known state be made to aid in the forensic investigation before powering down or shutting

Table 3
Highlights of the description, and result of challenges with provenance in cloud forensics.

Title of the challenge	Description	Challenge	Outcome of resolving challenge
The connection of evidence	It is difficult to keep track of activity across many cloud providers. Lack of interoperability is a major contributing factor.	Data from many cloud service providers must shows a connection.	When this issue of connection of evidence could be resolved, the investigator would be able to compare evidence from various cloud providers.
Creation of virtual storage	Cloud-based media imaging can be more complicated than it needs to be in some cases, and this might result in damage to the original material.	The creation of virtual storage in cloud settings via physical disk images.	Forensics investigators might more easily create trustworthy evidence for data stored on drives if this challenge could be addressed by reconstructing virtual storage from all key physical disk images.
The synchronization of time stamps	In a cloud system, time synchronization is much more difficult because timestamps must be synced across several physical servers dispersed over multiple geographical areas between both the cloud platform and distant web clients, including several end points.	Time stamp synchronization	If this challenge were overcome, it would be possible to achieve timestamp synchronization across the cloud environment relevant to the investigation (as may be needed for, e.g., correlation of evidence and timeline determination).
Deleting in the cloud	Deletion in the cloud is often based on the deletion of nodes pointing to information in virtual instances. Pathways for retrieval of the deleted information are dependent on cloud Providers offering sufficiently sophisticated mechanisms for access. Once the data is recovered, it remains a challenge to attribute specific data items to an individual user given the fact that cloud-based storage is a shared service in a multi-tenant environment.	Data recovery from the cloud that has been erased by either the adversary, consumer or cloud provider.	When this problem being solved, it would be a lot easier for data recovery and identification to take place.
Recovering deleted or corrupted files	It is impossible to recover data that has been rewritten by some other users in a sharing virtual environment, because the nodes referring to it have been erased. Users may overwrite their own data as well as that of other users. If the latter is the case, determining who owns what is a challenge.	Recovering of deleted data or lost files in a sharing virtual environment that has been modified by some other user	Deleted data that has been altered or corrupted can be recovered and associated with a specific user if this obstacle is overcome.
Incorporation of meta data	Since common fields such date of creation, date of last modification, date of last accessed, who create and who access etc., may change as the data is transferred to and within the cloud. The use of metadata such as an authentication technique could be at risk. Additionally, metadata can be changed while being collected, creating problems with spoliation worries and authentication. Entities that store information in the cloud should think about how the cloud will affect metadata, be aware of the metadata the cloud provider keeps, and determine whether it is easily accessible for e-discovery needs. However, cloud computing can provide more metadata that is not accessible in non-cloud situations such as versioning information, login events to the cloud environment, and file integrity data.	Metadata utilization	An investigator may have access to persisting metadata fields. However, if this problem could be overcome as may be required to like, establish timelines, highlight usage patterns event relation, and point to gaps in the data and events. Authentication may not be necessary in this challenge.
Identification of the malicious intrusion	Cyberattacks are often conducted in a series of small, gradual steps, where each exploits a seemingly insignificant weakness. The cloud space can likewise be exploited using this step-by-step strategy. There will not be a single "ah-ha" event where an attack is mounted and a system is breached, according to forensic investigators. However, they will probably discover a number of minute adjustments made to multiple different systems and programs that let an intruder compromise the intended destination.	Discovery of the criminal behavior	The detection of malicious behavior would be simpler and quicker if this obstacle could be resolved.

down a computer. The suspect's computer may have had critical information, but it would be lost once it was turned off because of its volatile nature. This potentially misleading data may still guide a researcher in the preliminary phases and provide a basis for further investigation even if it is ultimately not decisive. Therefore, it is important to initially collect the most easily lost data during a computer forensics examination.

Solutions to losing volatile data.

When using a mobile device, it is important to keep the battery always charged so that one does not lose any of the device's volatile storage. The only way to prevent the loss of evidence stored in volatile memory is to keep the computer system powered up until a forensics specialist can retrieve the data. There is a limited window of opportunity after closing the browser in which web-mail

(for example, Hotmail, as opposed to email programmes like Outlook), MSN Chat, etc., might be recoverable for forensic analysis. However, this is not always the case. Typically, the circumstances will determine the outcome.

Tools like The Sleuth Kit focus on the hard drive, but this is not the only place where forensic data and artifacts can be stored on a machine. Another good solution to prevent the loss of evidence stored in volatile memory is to apply volatility. Volatility is the most well-known and popular tool for the analysis of volatile memory. Volatility is free, open-source and supports third-party plugins. In fact, the Volatility Foundation holds an annual contest for users to develop the most useful and innovative extensions to the framework. Wireshark is another tool can be utilized to address loss of evidence stored in volatile memory. Wireshark is the go-to software for analyzing network traffic. Wireshark provides a clear and simple graphical user interface (GUI) for traffic analysis and a wealth of capabilities behind the scenes, plus it is free and open source. It may capture network traffic in real time or read previously recorded packets for analysis. Other tools that can be used are the computer-aided investigative environment (CAINE) and third-party plugin tools like Autopsy.

More also, a practical security operation center and existing forensic tools and techniques can address the issue of volatile data being overwritten. New blockchain technology can be introduced since it gives a promising decentralized platform to build tamper-proof systems. Its incorruptible distributed ledger/blockchain complements the need to maintain cloud data provenance. For instance, research by Shetty et al. (2017) proposed a blockchain-based cloud-based data provenance framework that tracks the operations of data records and produces provenance data. Furthermore, there is a need to dig more into the questions surrounding cloud computing provenance, especially as they pertain to the TClouds project's focus on areas like healthcare and public lighting. Both a top-down approach, detailing how a new provenance architecture may be crafted for clouds of various complexity, and a bottom-up one, wherein preexisting logging systems are integrated to meet requirements without disrupting preexisting ones, will need to be attempted. It is also possible that provenance systems for grid workflows can be adapted to keep track of conversations happening in the cloud.

18. Data provenance research gaps in cloud forensics

Data security is the main issue with cloud computing. Many cloud computing researchers have developed a considerable number of approaches for data protection, as was already covered in section 8. However, there are still some gaps in those techniques. This section focuses on how those gaps can be addressed. These issues must be resolved prior to using a complete provenance technique for cloud forensics. Therefore, the open questions and research gaps are discussed as follows:

(i) Confidentiality: Investigators may be unable to make knowledgeable decisions on network security statuses because of varied access control and privacy policy restrictions. On the other hand, partial exposure could result in inaccurate assumptions being made about the origin of data objects. Therefore, methods that provide an accurate perspective of provenance while maintaining security, privacy, and confidentiality are needed. Ruan et al. (2011) recognized the requirement for policies and standards as a path forward. However, in such policies and standards, the confidentiality of data provenance should be examined and considered. Therefore, more research on the confidentiality of data provenance in cloud computing is needed.

(ii) Outside-of-system provenance tracking: Users move data in and out of the cloud storage environment. This calls for the creation of techniques for tracing the origin of data objects even after they have left the confines of the system. Such methods are crucial for data accountability at a time when data is changing. Tan et al. (2013) proposed a prototype for tracking data traveling outside of a cloud system in order to address this problem. The approach encapsulates data in a self-executing container before leaving a cloud system. Consequently, the only way for investigators to access the data is via launching the viewer application that logs user actions and sends them to the cloud for provenance gathering while running the self-executing container. However, the integrity of the data that has been gathered could potentially be jeopardized if the data tracker is compromised. Because trusted computing is computationally expensive, practical alternatives are still needed. Trusted computing is one method for guaranteeing the integrity of the remote system where the data tracker is executing. Therefore, one of the weaknesses of data provenance in cloud computing is the lack of a multi-layer access policy.

(iii) Striking a balance between provenance-aware systems and discrepancy systems: It is often necessary to handle the problem of extracting provenance from the data itself or from incomplete provenance information since provenance information is frequently only partially or totally available. Integrating next-generation provenance-aware systems into current frameworks is one strategy. A technique for modifying an existing data system to facilitate provenance collection, storage, dissemination, and diffusion was reported by Conover et al. (2013). This approach automatically logs data inputs and transformations but manually compiles context data for auxiliary files. To enable interactive user access, provenance and context data are integrated and incorporated into XML documents. This technique requires a lot of processing power; thus, it is still being researched and needs more researchers to look into it.

(iv) Requires trust establishment in Third Party Auditor (TPA) for dispute resolution: Although cloud computing has clearly advanced, users still have some reservations about its level of security and adoption-related problems that keep them from fully trusting this practical technology. Therefore, a number of security paradigms for cloud computing based on a third-party auditor (TPA) have been established in order to strengthen the confidence between cloud clients (CC), cloud service providers (CSP) and to protect the CC's data in the cloud. The integrity of the CC's data and any pertinent information connected to it must be verified by the TPA as a trusted party (Razaque et al., 2021). However, the TPA can turn against the CC and try to compromise the data's privacy by acting maliciously.

19. Future directions

Future work on provenance security and privacy should focus on many vital areas.

1. The focus of the current study on provenance security has been on identifying needs and proposing solutions utilizing tried-and-true methods for data security. It would be intriguing to learn whether, like with privacy, the interaction between data and provenance leads to new security issues and solutions.
2. There is a disconnect in this field between theory and practice. For instance, PROV, a standard for provenance representation and exchange, has been proposed and is beginning to be imple-

mented in provenance-enabled applications. However, the methods that permit provenance privacy and security lack some abstractions, and concepts such as composite modules and hybrids are significant for mechanisms that provide provenance security and privacy. More so, the lack of actual datasets and well-specified measures on availability and confidentiality needs has prevented satisfactory evaluations of the practical efficiency of provenance privacy strategies. It is critical to close this gap in upcoming research.

3. The privacy concepts that are currently being employed for provenance are somewhat flimsy and rely heavily on presumptions. For instance, the privacy module employs diversity (Machanavajjhala et al., 2007). However, assuming the opponent has no prior knowledge, it is accomplished by concealing some provenance information via authentication protocols. Nonetheless, there are considerably more robust ideas of privacy, including differential privacy (Davidson & Roy, 2018; Dwork, 2008), that depend on introducing random noise to the results of aggregate searches. Understanding how these methods might be applied in the provenance setting, where queries are frequently not aggregated and repeatability is crucial, is critical.

20. Conclusion

Recently, cloud computing has grown in popularity, with high efficiency, reliability, and low cost. Similarly, cyber-attacks on the cloud have increased probably due to poverty, lack of job and moral decadence in society. Interestingly, digital forensics is widely adopted to investigate cyber-attacks scenario to provide evidence that can be used to prosecute the offender. Although, performing digital forensic investigations in the virtual world is a current and pressing matter. However, digital forensics come with enormous challenges with the revolution of the Internet of Things (IoTs) where heterogeneous systems are interconnected, thus, providing intelligent and advanced services for human benefits and industrial production.

Due to its many advantages and bright future, cloud computing has attracted a plethora of research attention. The affordable nature of its services has helped it become more well-liked among businesses. Although cloud computing offers appealing service packages for the commercial sector, its acceptance is still quite gradual, primarily because of the security and privacy issues that are present in the industry. An in-depth analysis of various data provenance challenges in the cloud for digital forensics, such as in wireless sensor networks, IoT, and blockchain, has been elaborated. This article also examined the three layers of cloud architecture's three current provenance-collection methodologies, noting the shortcomings of each method. Provenance collection issues are described and analyzed, as well as the bare needs for efficient cloud forensics provenance collection.

The analysis of current cloud provenance gathering methods and the identification of crucial areas requiring additional and fresh research constitute the paper's key contributions. Confidentiality, tracking provenance outside of a system, and bridging gaps between current and provenance-aware systems are the three areas where gaps have been found. In the absence of conventional "seize and capture" methods, addressing these key areas can give cloud forensic investigators a mechanism to gather investigatory provenance evidence. Furthermore, the creation of improved provenance solutions will result from efficiently addressing these shortcomings or difficulties. As the importance of reliable electronic record-keeping increases, assurances on the security of provenance information are more crucial than ever.

The collection and examination of digital evidence may become more challenging due to cloud computing, as so many computer devices in the cloud could require forensic analysis. Cloud computing system computer forensic investigations will likely require more time and effort. In addition, it can be challenging to determine what data was stored or processed on what particular computing device by what program. Computer forensic investigators may find it more challenging to gather and examine digital evidence to the same standards currently expected for conventional server-based systems as a result of cloud computing systems. The future plan will explore the prerequisites for cloud computing provenance, especially in the context of the TClouds project's applications like healthcare and public lighting. The TClouds project focuses on cross-border infrastructure privacy protection and resilience against failures and assaults to address increasing concerns about the security of personal data in the cloud 6789.

Funding

This research is funded by the Office of Research & Sponsored Programs of Abu Dhabi University, United Arab Emirates and this work has been partially supported by Abu Dhabi University under Grant No 19300635.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abbadi, I.M., 2011. Toward trustworthy clouds' internet scale critical infrastructure. In: International Conference on Information Security Practice and Experience. Springer, Berlin, Heidelberg, pp. 71–82.

Abbadi, I. M., & Lyle, J. (2011). Challenges for provenance in cloud computing. In 3rd USENIX Workshop on the Theory and Practice of Provenance (TaPP 11).

Abbott, L., 2018. The utility of offensive cyber-operations in conventional military engagements. Australian Army Journal 14 (2), 5–14.

AbdElnapi, N.M., Omara, F.A., Omran, N.F., 2016. A hybrid hashing security algorithm for data storage on cloud computing. International Journal of Computer Science and Information Security (IJCSIS) 14 (4).

Abernathey, R.P., Augspurger, T., Banihirwe, A., Blackmon-Luca, C.C., Crone, T.J., Gentemann, C.L., Signell, R.P., 2021. Cloud-native repositories for big scientific data. Computing in Science & Engineering 23 (2).

Abiodun, E.O., Alabdulatif, A., Abiodun, O.I., Alawida, M., Alabdulatif, A., Alkhawaldeh, R.S., 2021a. A systematic review of emerging feature selection optimization methods for optimal text classification: the present state and prospective opportunities. Neural Computing and Applications 33 (22), 15091–15118.

Abiodun, O.I., Abiodun, E.O., Alawida, M., Alkhawaldeh, R.S., Arshad, H., 2021b. A review on the security of the internet of things: challenges and solutions. Wireless Personal Communications 119 (3), 2603–2637.

Abiodun, O. I., Jantan, A., Omolara, A. E., Mahinderjit, M. M., Abubakar, Z. L., & Umar, A. M. (2018). Big Data: an approach for detecting terrorist activities with people's profiling. In proceedings of the International MultiConference of Engineers and Computer Scientists (Vol. 1).

Abiodun, O.I., Jantan, A., Singh, M.M., Anbar, M., Zaaba, Z.F., Abiodun, O.E.O., 2018b. Forensic DNA profiling for identifying an individual crime. International Journal of Civil Engineering and Technology 9 (7), 755–765.

Act, A., 1996. Health insurance portability and accountability act of 1996. Public law 104, 191.

Agrawal, P., Benjelloun, O., Sarma, A. D., Hayworth, C., Nabar, S., Sugihara, T., & Widom, J. (2006). Trio: A system for data, uncertainty, and lineage. Proc. of VLDB 2006 (demonstration description).

Ahmed, M., Chowdhury, A.S.M.R., Ahmed, M., Rafee, M.M.H., 2012. An advanced survey on cloud computing and state-of-the-art research issues. IJCSI International Journal of Computer Science Issues 9 (1).

Akin, O.C., Matthew, F., Comfort, D., 2014. The impact and challenges of cloud computing adoption on public universities in Southwestern Nigeria. International Journal of Advanced Computer Science and Applications (IJACSA) 5 (8), 13–19.

Alawida, M., Omolara, A. E., Abiodun, O. I., & Al-Rajab, M. (2022). A deeper look into cybersecurity issues in the wake of Covid-19: a survey. Journal of King Saud University-Computer and Information Sciences.

Alhomdy, S., Thabit, F., Abdulrazzak, F.A.H., Haldorai, A., Jagtap, S., 2021. The role of cloud computing technology: A savior to fight the lockdown in COVID 19 crisis, the benefits, characteristics and applications. International Journal of Intelligent Networks 2, 166–174.

Al-Hujran, O., Al-Lozi, E.M., Al-Debei, M.M., Maqableh, M., 2018. Challenges of cloud computing adoption from the TOE framework perspective. International Journal of E-Business Research (IJEBR) 14 (3), 77–94.
Alkhalil, A., Ramadan, R.A., 2017. IoT data provenance implementation challenges. Procedia Computer Science 109, 1134–1139.
Al-Khateeb, H., Epiphaniou, G., Daly, H., 2019. Blockchain for modern digital forensics: The chain-of-custody as a distributed ledger. In: Blockchain and Clinical Trial. Springer, Cham, pp. 149–168.
Alnabelsi, S. H., Salameh, H. B., Saifan, R. R., & Darabkh, K. A. (2022). A multi-layer hyper-graph routing with jamming-awareness for improved throughput in full-duplex cognitive radio networks. Journal of King Saud University-Computer and Information Sciences.
Alqahtany, S., Clarke, N., Furnell, S., Reich, C., 2015. Cloud forensics: a review of challenges, solutions and open problems. In: In 2015 international conference on cloud computing (ICCC). IEEE, pp. 1–9.
Alruwaili, F.F., 2021. CustodyBlock: A Distributed Chain of Custody Evidence Framework. Information 12 (2), 88.
Aneja, M.J.S., Bhatia, T., Sharma, G., Shrivastava, G., 2018. In: Artificial intelligence based intrusion detection system to detect flooding attack in VANETs. IGI Global, pp. 87–100.
Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R., Konwinski, A., Zaharia, M., 2010. A view of cloud computing. Communications of the ACM 53 (4), 50–58.
Arshad, H., Omlara, E., Abiodun, I.O., Aminu, A., 2020. A semi-automated forensic investigation model for online social networks. Computers & Security 97, 101946.
Arshad, H., Abdullah, S., Alawida, M., Alabdulatif, A., Abiodun, O.I., Riaz, O., 2022. A Multi-Layer Semantic Approach for Digital Forensics Automation for Online Social Networks. Sensors 22 (3), 1115.
As'habi, K., Vafabakhsh, A., Borji, S., 2016. Data transmission security in clouds computing. Indian Journal of Fundamental and Applied Life Sciences 6, 2231–6345.
Asghar, M.R., Ion, M., Russello, G., Crispo, B., 2011. In: June). Securing data provenance in the cloud. Springer, Berlin, Heidelberg, pp. 145–160.
Ayers, R. P., Brothers, S., & Jansen, W. (2014). Guidelines on Mobile Device Forensics. (National Institute of Standards and Technology, Gaithersburg, MD). NIST Special Publication (SP). 800-101.
Badharudheen, P., Chacko, A.M., Madhu Kumar, S.D., 2014. Making an Application Provenance-Aware through UML–A General Scheme. In: International Conference on Security in Computer Networks and Distributed Systems. Springer, Berlin, Heidelberg, pp. 451–460.
Baesens, B., 2014. Analytics in a big data world: The essential guide to data science and its applications. John Wiley & Sons.
Baracaldo, N., Bathen, L.A.D., Ozugha, R.O., Engel, R., Tata, S., Ludwig, H., 2016. Securing data provenance in internet of things (IoT) systems. In: International conference on service-oriented computing. Springer, Cham, pp. 92–98.
Barga, R.S., Digiampietri, L.A., 2008. Automatic capture and efficient storage of e-Science experiment provenance. Concurrency and Computation: Practice and Experience 20 (5), 419–429.
Bates, A., Mood, B., Valafar, M., & Butler, K. (2013, February). Towards secure provenance-based access control in cloud environments. In Proceedings of the third ACM conference on Data and application security and privacy (pp. 277-284).
Bayramusta, M., Nasir, V.A., 2016. A fad or future of IT?: A comprehensive literature review on the cloud computing research. International Journal of Information Management 36 (4), 635–644.
Bell, M.J., Collison, M., Lord, P., 2013 Oct 15. Can inferred provenance and its visualisation be used to detect erroneous annotation? A case study using UniProtKB. PLoS One 8 (10), e75541.
Bem, D., Huebner, E., 2007. Computer forensic analysis in a virtual environment. International journal of digital evidence 6 (2), 1–13.
Bertino, E., 2014. Data trustworthiness—approaches and research challenges. Data privacy management, autonomous spontaneous security, and security assurance, 17–25.
Bertino, (2016). Data Security and Privacy in the IoT, in Proceedings of the 19th International Conference on Extending Database Technology, 2016, pp. 1–3.
Bhisikar, P., Sahu, A., 2013. Security in data storage and transmission in cloud computing. International journal of advanced research in computer science and software engineering 3 (3).
Birk, D., Wegener, C., 2011. Technical issues of forensic investigations in cloud computing environments. In: In 2011 Sixth IEEE international workshop on systematic approaches to digital forensic engineering. IEEE, pp. 1–10.
Booth, D., Haas, H., McCabe, F., Newcomer, E., Champion, M., FER-RIS, C., & OR-CHARD, D. (2004). Web Service Architecture http://www. w3. org. TR/ws-arch.
Bowers, S., McPhillips, T.M., Ludäscher, B., 2008. Provenance in collection-oriented scientific workflows. Concurrency and Computation: Practice and Experience 20 (5), 519–529.
Braun, U., Garfinkel, S., Holland, D.A., Muniswamy-Reddy, K.K., Seltzer, M.I., 2006. Issues in automatic provenance collection. In: International Provenance and Annotation Workshop. Springer, Berlin, Heidelberg, pp. 171–183.
Braun, U.J., Shinnar, A., Seltzer, M.I., 2008. Securing provenance. Usenix Association.
Buneman, P., Chapman, A., & Cheney, J. (2006, June). Provenance management in curated databases. In Proceedings of the 2006 ACM SIGMOD international conference on Management of data, pp. 539–550.
Buneman, P., Tan, W.C., 2019. Data provenance: What next? ACM SIGMOD Record 47 (3), 5–16.
Cadenhead, T., Khadilkar, V., Kantarcioglu, M., & Thuraisingham, B. (2011, February). A language for provenance access control. In Proceedings of the first ACM conference on Data and application security and privacy (pp. 133-144).
Casey, E., 2011. Digital evidence and computer crime: Forensic science, computers, and the internet. Academic press.
Cassim, Z., Handjiski, B., Schubert, J., Zouaoui, Y., 2020. The $10 trillion rescue: How governments can deliver impact. McKinsey & Company.
Cebe, M., Erdin, E., Akkaya, K., Aksu, H., Uluagac, S., 2018. Block4forensic: An integrated lightweight blockchain framework for forensics applications of connected vehicles. IEEE communications magazine 56 (10), 50–57.
Chakraborti, A., Curtmola, R., Katz, J., Nieh, J., Sadeghi, A.R., Sion, R., Zhang, Y., 2022. Cloud Computing Security: Foundations and Research Directions. Foundations and Trends® in Privacy and Security 3 (2), 103–213.
Chan, S. C., Cheney, J., Bhatotia, P., Pasquier, T., Gehani, A., Irshad, H., ... & Seltzer, M. (2019, December). ProvMark: a provenance expressiveness benchmarking system. In Proceedings of the 20th International Middleware Conference (pp. 268-279).
Chapman, A., Missier, P., Simonelli, G., Torlone, R., 2020. Capturing and querying fine-grained provenance of preprocessing pipelines in data science. Proceedings of the VLDB Endowment 14 (4), 507–520.
Chavan, A., Jadhav, A., Kumbhar, S., Joshi, I., 2019. Data transmission using RSA algorithm. International Research Journal of Engineerin and Technology 6 (3), 34–36.
Cheah, Y.W., Canon, R., Plale, B., Ramakrishnan, L., 2013. Milieu: Lightweight and configurable big data provenance for science. In: In 2013 IEEE International Congress on Big Data. IEEE, pp. 46–53.
Chebotko, A., Abraham, J., Brazier, P., Piazza, A., Kashlev, A., Lu, S., 2013. Storing, indexing and querying large provenance data sets as RDF graphs in apache HBase. In: In 2013 IEEE Ninth World Congress on Services. IEEE, pp. 1–8.
Chebotko, A., Chang, S., Lu, S., Fotouhi, F., & Yang, P. (2010). Secure scientific workflow provenance querying with security views. In Proceedings of the Ninth International Conference on Web-Age Information Management (pp. 349-356).
Chen, X., Zeng, D., Pang, S., & Jun, F. (2021). Cloud Computing Storage Data Access Control Method Based on Dynamic Re-Encryption. Security and Communication Networks, 2021.
Cheney, J., Chong, S., Foster, N., Seltzer, M., & Vansummeren, S. (2009, October). Provenance: a future history. In Proceedings of the 24th ACM SIGPLAN conference companion on Object oriented programming systems languages and applications (pp. 957-964).
Cheney, J., Chiticariu, L., Tan, W.C., 2009b. Provenance in Databases: Why, How, and where. Now Publishers Inc..
Codd, E.F., 1970. A relational model of data for large shared data banks. Communications of the ACM 13 (6), 377–387.
Conover, H., Ramachandran, R., Beaumont, B., Kulkarni, A., McEniry, M., Regner, K., Graves, S., 2013. Introducing provenance capture into a legacy data system. IEEE transactions on geoscience and remote sensing 51 (11), 5098–5104.
Cook, A., Robinson, M., Ferrag, M.A., Maglaras, L.A., He, Y., Jones, K., Janicke, H., 2018. Internet of cloud: Security and privacy issues. In: Cloud Computing for Optimization: Foundations, Applications, and Challenges. Springer, Cham, pp. 271–301.
Cooper, J., James, A., 2009. Challenges for database management in the internet of things. IETE Technical Review 26 (5), 320–329.
Cosic, J., Cosic, Z., 2012. Chain of custody and life cycle of digital evidence. Computer technology and application 3 (2).
Crawl, D., Altintas, I., 2008. A provenance-based fault tolerance mechanism for scientific workflows. In: International Provenance and Annotation Workshop. Springer, Berlin, Heidelberg, pp. 152–159.
Cruz, S.M., Campos, M.L., Mattoso, M., 2009 Jul 6. Towards a taxonomy of provenance in scientific workflow management systems. In: Services-I, 2009 World Conference on. IEEE, pp. 259–266.
Cui, Y., Widom, J., 2003. Lineage tracing for general data warehouse transformations. VLDB J. Int. J. Very Larg. Data 12 (1), 41–58.
Cummings, D.I., Dumas, S., Dalrymple, R.W., 2009. Fine-grained versus coarse-grained wave ripples generated experimentally under large-scale oscillatory flow. Journal of Sedimentary Research 79 (2), 83–93.
Curcin, V. (2017). Embedding data provenance into the Learning Health System to facilitate reproducible research (Vol. 1, No. 2, p. e10019). Chichester, UK: John Wiley & Sons, Ltd.
da Silva, P.P., McGuinness, D.L., McCool, R., 2003. Knowledge provenance infrastructure. 26 (4), 26–32.
Davidson, S. B., & Roy, S. (2018). Provenance: Privacy and Security.
Davidson, S.B., Freire, J., 2008. Jun 9. Provenance and scientific workflows: challenges and opportunities. In: In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. ACM, pp. 1345–1350.
De Kinderen, S., Kaczmarek-Heß, M., Ma, Q., Razo-Zapata, I.S., 2017. Towards Meta Model Provenance: A Goal-Driven Approach to Document the Provenance of Meta Models. In: In 10th IFIP Working Conference on The Practice of Enterprise Modeling (PoEM). Springer International Publishing, pp. 49–64.
Deepa, N., Pham, Q.V., Nguyen, D.C., Bhattacharya, S., Prabadevi, B., Gadekallu, T.R., Pathirana, P.N., 2022. A survey on blockchain for big data: approaches, opportunities, and future directions. Future Generation Computer Systems.
Ding, W., Jing, X., Yan, Z., Yang, L.T. (2019). A survey on data fusion in internet of things: towards secure and privacy-preserving fusion. Information Fusion. 51, 129 –144 (2019).

Dogan, G., 2016. A Survey of Provenance in Wireless Sensor Networks. Adhoc & Sensor. Wireless Networks 30.
Drăgoi, C., Henzinger, T.A., Zufferey, D., 2016. PSync: a partially synchronous language for fault-tolerant distributed algorithms. ACM SIGPLAN Notices 51 (1), 400–415.
Dwork, C., 2008. Differential privacy: A survey of results. In: International conference on theory and applications of models of computation. Springer, Berlin, Heidelberg, pp. 1–19.
Elkhodr, M., & Mufti, Z. B. (2019). On the challenges of data provenance in the Internet of Things. arXiv preprint arXiv:1907.07316.
Farahani, B., Firouzi, F., Luecking, M., 2021. The convergence of IoT and distributed ledger technologies (DLT): Opportunities, challenges, and solutions. Journal of Network and Computer Applications 177, 102936.
Foster, I. T., Vöckler, J. S., Wilde, M., & Zhao, Y. (2003, January). The Virtual Data Grid: A New Model and Architecture for Data-Intensive Collaboration. In CIDR.
Foster, I., Kesselman, C., Tuecke, S., 2001. The anatomy of the grid: Enabling scalable virtual organizations. The International Journal of High Performance Computing Applications 15 (3), 200–222.
Freire, J., Koop, D., Santos, E., Silva, C.T., 2008. Provenance for computational tasks: A survey. Computing in science & engineering 10 (3), 11–21.
Fromholz, H.J., 1977. Discovery, Evidence, Confidentiality, and Security Problems Associated with the Use of Computer-Based Litigation Support Systems. Wash, ULQ, p. 445.
Gadelha Jr, L.M., Clifford, B., Mattoso, M., Wilde, M., Foster, I., 2011. Provenance management in Swift. Future Generation Computer Systems 27 (6), 775–780.
Gadelha Jr, L.M., Mattoso, M., 2008. Kairos: an architecture for securing authorship and temporal information of provenance data in grid-enabled workflow management systems. In: In 2008 IEEE Fourth International Conference on eScience. IEEE, pp. 597–602.
Galhardas, H., Florescu, D., Shasha, D. E., Simon, E., & Saita, C. A. (2001, June). Improving Data Cleaning Quality Using a Data Lineage Facility. In DMDW (p. 3).
Gauttam, H., Pattanaik, K.K., Bhadauria, S., Saxena, D., 2022. A cost aware topology formation scheme for latency sensitive applications in edge infrastructure-as-a-service paradigm. Journal of Network and Computer Applications 199, 103303.
Gazis, V., Goertz, M., Huber, M., Leonardi, A., Mathioudakis, K., Wiesmaier, A., Zeiger, F., 2015. Short paper: IoT: Challenges, projects, architectures. In: In 2015 18th international conference on intelligence in next generation networks. IEEE, pp. 145–147.
Gehani, A., Kim, M., Zhang, J., (2009 Feb 23). Steps toward managing lineage metadata in grid clusters. In: First Workshop on on Theory and Practice of Provenance. USENIX Association, p. 7.
Gehani, A., Lindqvist, U., 2007. Veil: A system for certifying video provenance. In: Ninth IEEE International Symposium on Multimedia (ISM 2007). IEEE, pp. 263–272.
George, M., Mary Chacko, A., 2022. MediTrans—Patient-centric interoperability through blockchain. International Journal of Network Management 32 (3), e2187.
Gercke, M. (2016). Understanding cybercrime: a guide for developing countries.
Ghoshal, D., & Plale, B. (2013, March). Provenance from log files: a BigData problem. In Proceedings of the Joint EDBT/ICDT 2013 Workshops (pp. 290-297).
Giova, G., 2011. Improving chain of custody in forensic investigation of electronic digital systems. International Journal of Computer Science and Network Security 11 (1), 1–9.
Glavic, B., 2012. Big data provenance: Challenges and implications for benchmarking. Specifying big data benchmarks, 72–80.
Glavic, B., Esmaili, K. S., Fischer, P. M., & Tatbul, N. (2011). The case for fine-grained stream provenance.
Global guidelines, (2019). INTERPOL. Global guidelines for digital forensics laboratories.
Goble, C. (2002, October). Position statement: Musings on provenance, workflow and (semantic web) annotations for bioinformatics. In Workshop on Data Derivation and Provenance, Chicago (Vol. 3).
Golbeck, J., Hendler, J., 2008. A semantic web approach to the provenance challenge. Concurrency and Computation: Practice and Experience 20 (5), 431–439.
Goldwasser, S., Micali, S., Rivest, R.L., 1988. A digital signature scheme secure against adaptive chosen-message attacks. SIAM Journal on computing 17 (2), 281–308.
Greenwood, M., Goble, C., Stevens, R., Zhao, J., Addis, M., Marvin, D., ... & Oinn, T. (2003). Provenance of e-science experiments-experience from bioinformatics.
Griggs, K.N., Ossipova, O., Kohlios, C.P., Baccarini, A.N., Howson, E.A., Hayajneh, T., 2018. Healthcare blockchain system using smart contracts for secure automated remote patient monitoring. Journal of medical systems 42 (7), 1–7.
Grispos, G., Storer, T., Glisson, W.B., 2012. Calm before the storm: The challenges of cloud computing in digital forensics. International Journal of Digital Crime and Forensics (IJDCF) 4 (2), 28–48.
Guo, H., Jin, B., Huang, D., 2010. Research and review on computer forensics. In: International Conference on Forensics in Telecommunications, Information, and Multimedia. Springer, Berlin, Heidelberg, pp. 224–233.
Gupta, M., Sharman, R., 2012. Determinants of data breaches: A categorization-based empirical investigation. Journal of Applied Security Research 7 (3), 375–395.
Haque, S., Atkison, T., 2018. A forensic enabled data provenance model for public cloud. Journal of Digital Forensics, Security and Law 13 (3), 7.
Hasan, R., Sion, R., & Winslett, M. (2007, October). Introducing secure provenance: problems and challenges. In Proceedings of the 2007 ACM workshop on Storage security and survivability (pp. 13-18).
Hasan, R., Sion, R., Winslett, M., 2009. Preventing history forgery with secure provenance. ACM Transactions on Storage (TOS) 5 (4), 1–43.
Hastings, J., Chepelev, L., Willighagen, E., Adams, N., Steinbeck, C., Dumontier, M., 2011. The chemical information ontology: provenance and disambiguation for chemical data on the biological semantic web. PLoS One 6 (10), e25513.
Herman, M., Iorga, M., Salim, A.M., Jackson, R.H., Hurst, M.R., Leo, R., Sardinas, R., 2020. Nist cloud computing forensic science challenges. National Institute of Standards and Technology, Gaithersburg, Maryland.
Herschel, M., Diestelkämper, R., Ben Lahmar, H., 2017. A survey on provenance: What for? What form? What from? The VLDB Journal 26 (6), 881–906.
Hoefer, C.N., Karagiannis, G., 2010. In: December). Taxonomy of cloud computing services. IEEE, pp. 1345–1350.
Howgrave-Graham, N.A., Smart, N.P., 2001. Lattice attacks on digital signature schemes. Designs, Codes and Cryptography 23 (3), 283–290.
Hu, R., Yan, Z., Ding, W., Yang, L.T., 2020. A survey on data provenance in IoT. World Wide Web 23 (2), 1441–1463.
Huang, K., Siegel, M., Madnick, S., 2018. Systematically understanding the cyber attack business: A survey. ACM Computing Surveys (CSUR) 51 (4), 1–36.
Ieong, R.S., 2006. FORZA–Digital forensics investigation framework that incorporate legal issues. digital investigation 3, 29–36.
Iftekhar, A., Cui, X., Hassan, M., & Afzal, W. (2020). Application of blockchain and Internet of Things to ensure tamper-proof data availability for food safety. Journal of Food Quality, 2020.
Ikeda, R., & Widom, J. (2010). Panda: A system for provenance and data.
Iturbide, M., Bedia, J., Herrera, S., Baño-Medina, J., Fernández, J., Frías, M.D., Gutiérrez, J.M., 2019. The R-based climate4R open framework for reproducible climate data access and post-processing. Environmental Modelling & Software 111, 42–54.
Jacobson, M.H., 1977. The Use of Computer Printouts as Evidence in Commercial Litigation. Com. LJ 82, 14.
Jagadish, H.V., Olken, F., 2004. Database management for life sciences research. ACM SIGMOD Record 33 (2), 15–20.
Jain, A., Kumar, R., 2014. A taxonomy of cloud computing. International journal of scientific and research publications 4 (7), 1–5.
Jangjou, M., Sohrabi, M.K., 2022. A comprehensive survey on security challenges in different network layers in cloud computing. Archives of Computational Methods in Engineering, 1–22.
Janssen, M., van den Hoven, J., 2015. Big and Open Linked Data (BOLD) in government: A challenge to transparency and privacy? Government Information Quarterly 32 (4), 363–368.
Jennath, H. S., Anoop, V. S., & Asharaf, S. (2020). Blockchain for healthcare: securing patient data and enabling trusted artificial intelligence.
Jensen, M., Gruschka, N., & Luttenberger, N. (2008, March). The impact of flooding attacks on network-based services. In 2008 Third International Conference on Availability, Reliability and Security (pp. 509-513). IEEE.
Kahn, M.G., Mui, J.Y., Ames, M.J., Yamsani, A.K., Pozdeyev, N., Rafaels, N., Brooks, I. M., 2022. Migrating a research data warehouse to a public cloud: challenges and opportunities. Journal of the American Medical Informatics Association 29 (4), 592–600.
Kangarlou, A., Eugster, P., Xu, D., 2011. VNsnap: Taking snapshots of virtual networked infrastructures in the cloud. IEEE Transactions on Services Computing 5 (4), 484–496.
Karvounarakis, G., Ives, Z. G., & Tannen, V. (2010, June). Querying data provenance. In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data (pp. 951-962).
Karyda, M., Mitrou, L., 2007. Internet forensics: Legal and technical issues. In: Second International Workshop on Digital Forensics and Incident Analysis (WDFIA 2007). IEEE, pp. 3–12.
Kashliev, A., 2020. Storage and Querying of Large Provenance Graphs Using NoSQL DSE. In: In 2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS). IEEE, pp. 260–262.
Katilu, V.M., Franqueira, V.N., Angelopoulou, O., 2015. Challenges of data provenance for cloud forensic investigations. In: In 2015 10th International Conference on Availability, Reliability and Security. IEEE, pp. 312–317.
Kebande, V.R., Venter, H.S., 2018. On digital forensic readiness in the cloud using a distributed agent-based solution: issues and challenges. Australian Journal of Forensic Sciences 50 (2), 209–238.
Kifor, T., Varga, L.Z., Vázquez-Salceda, J., Alvarez, S., Willmott, S., Miles, S., Moreau, L., 2006. Provenance in agent-mediated healthcare systems. IEEE Intelligent Systems 21 (6), 38–46.
King, S. T., Mao, Z. M., Lucchetti, D. G., & Chen, P. M. (2005, February). Enriching Intrusion Alerts Through Multi-Host Causality. In Ndss.
Ko, R.K., Jagadpramana, P., Lee, B.S., 2011. Flogger: A file-centric logger for monitoring file access and transfers within cloud computing environments. In: In 2011IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications. IEEE, pp. 765–771.
Lange, R.J., 2010. Provenance aware sensor networks for real-time data analysis. University of Twente). Master's thesis.
Langmead, B., Nellore, A., 2018. Cloud computing for genomic data analysis and collaboration. Nature Reviews Genetics 19 (4), 208–219.
Laskey, K.J., 2005. Metadata concepts to support a net-centric data environment. In: Net-centric approaches to intelligence and national security. Springer, Boston, MA, pp. 29–54.

Lea, D., 2000. Concurrent programming in Java: design principles and patterns. Addison-Wesley. Professional.
Lee, L.S., Brink, W.D., 2020. Trust in cloud-based services: A framework for consumer adoption of software as a service. Journal of Information Systems 34 (2), 65–85.
Lewis-Williams, J.D., 2016. Myth and meaning: San-Bushman folklore in global context. Routledge.
Li, Z., Chen, Q.A., Yang, R., Chen, Y., Ruan, W., 2021. Threat detection and investigation with system-level provenance graphs: a survey. Computers & Security 106, 102282.
Liang, X., Shetty, S., Tosh, D., Kamhoua, C., Kwiat, K., Njilla, L., 2017. Provchain: A blockchain-based data provenance architecture in cloud environment with enhanced privacy and availability. In: In 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID). IEEE, pp. 468–477.
Liu, Y., Nakatsuka, Y., Sani, A. A., Agarwal, S., & Tsudik, G. (2022, June). Vronicle: verifiable provenance for videos from mobile devices. In Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services (pp. 565-566).
Liu, Q., Gu, J., Yang, J., Li, Y., Sha, D., Xu, M., Yang, C., 2021. Cloud, Edge, and Mobile Computing for Smart Cities. In: Urban Informatics. Springer, Singapore, pp. 757–795.
Loch, K.D., Carr, H.H., Warkentin, M.E., 1992. Threats to information systems: today's reality, yesterday's understanding. Mis Quarterly, 173–186.
Lone, A.H., Mir, R.N., 2019. Forensic-chain: Blockchain based digital forensics chain of custody with PoC in Hyperledger Composer. Digital investigation 28, 44–55.
Loo, B. T., Condie, T., Hellerstein, J. M., Maniatis, P., Roscoe, T., & Stoica, I. (2005, October). Implementing declarative overlays. In Proceedings of the twentieth ACM symposium on Operating systems principles (pp. 75-90).
Lu, R., Lin, X., Liang, X., & Shen, X. (2010, April). Secure provenance: the essential of bread and butter of data forensics in cloud computing. In Proceedings of the 5th ACM symposium on information, computer and communications security (pp. 282-292).
Lyle, J., Martin, A., 2010. January). Better together. Usenix, Trusted computing and provenance.
Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M., 2007. l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD) 1 (1), 3-es.
Macko, P., & Chiarini, M. (2011). Collecting provenance via the Xen hypervisor. In 3rd USENIX Workshop on the Theory and Practice of Provenance (TaPP 11).
Malaverri, J.E.G., Santanchè, A., Medeiros, C.B., 2014. A provenance–based approach to evaluate data quality in eScience. International Journal of Metadata, Semantics and Ontologies 9 (1), 15–28.
Manral, B., Somani, G., Choo, K.K.R., Conti, M., Gaur, M.S., 2019. A systematic survey on cloud forensics challenges, solutions, and future directions. ACM Computing Surveys (CSUR) 52 (6), 1–38.
Mars, D., Mettali Gammar, S., Lahmadi, A., Azouz Saidane, L., 2019. Using information centric networking in internet of things: a survey. Wireless Personal Communications 105 (1), 87–103.
Martini, B., Choo, K.K.R., 2014. In: September). Remote programmatic vCloud forensics: a six-step collection process and a proof of concept. IEEE, pp. 935–942.
Marturana, F., Me, G., Tacconi, S., 2012. A case study on digital forensics in the cloud. In: In 2012 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery. IEEE, pp. 111–116.
McDaniel, P., 2011. Data provenance and security. IEEE Security & Privacy 9 (2), 83–85.
McDowall, R.D., 2013. How Complete Are Your Data? Spectroscopy 28 (4), 18–25.
McGuire, M., 2019. It ain't what it is, it's the way that they do it? Why we still don't understand cybercrime. In: The human factor of cybercrime. Routledge, pp. 3–28.
Mell, P., & Grance, T. (2011). The NIST definition of cloud computing.
Miles, S., Groth, P., Branco, M., Moreau, L., 2007. The requirements of recording and using provenance in e-Science. Journal of Grid Computing 5 (1), 1–25.
Miranda Lopez, E., Moon, S.Y., Park, J.H., 2016. Scenario-based digital forensics challenges in cloud computing. Symmetry 8 (10), 107.
Moreau, L., 2010. The foundations for provenance on the web. Found. Trends Web Sci. 2 (2–3), 99–241.
Moreau, L., 2011. Provenance-based reproducibility in the semantic web. Journal of Web Semantics 9 (2), 202–221.
Moreau, L., & Ali, M. (2014). A provenance-based policy control framework for cloud services.
Mothukuri, V., Cheerla, S.S., Parizi, R.M., Zhang, Q., Choo, K.K.R., 2021. BlockHDFS: Blockchain-integrated Hadoop distributed file system for secure provenance traceability. Blockchain: Research and Applications 2, (4) 100032.
Mufti, Z. B., & Elkhodr, M. (2018). Data provenance in the internet of things: Views and challenges. CSEN, NCWC, 1-7.
Müller, T., Dietrich, B., & Grust, T. (2018). You Say'What', I Hear'Where'and'Why': (Mis-) Interpreting SQL to Derive Fine-Grained Provenance. arXiv preprint arXiv:1805.11517.
Mulugeta, L., Drach, A., Erdemir, A., Hunt, C.A., Horner, M., Ku, J.P., Lytton, W.W., 2018. Credibility, replicability, and reproducibility in simulation for biomedicine and clinical applications in neuroscience. Frontiers in Neuroinformatics 12, 18.
Muniswamy-Reddy, K. K., Macko, P., & Seltzer, M. I. (2009, February). Making a Cloud Provenance-Aware. In Workshop on the Theory and Practice of Provenance.
Muniswamy-Reddy, K.K., Seltzer, M., 2010. Provenance as first class cloud data. ACM SIGOPS Operating Systems Review 43 (4), 11–16.
Murphy, D., 2000. The Discovery of Electronic Data in Litigation: What Practitioners and Their Clients Need to Know. Wm. Mitchell L. Rev. 27, 1825.
Namasudra, S., Deka, G.C., Johri, P., Hosseinpour, M., Gandomi, A.H., 2021. The revolution of blockchain: State-of-the-art and research challenges. Archives of Computational Methods in Engineering 28 (3), 1497–1515.
No, U. P. L. (2002). 107-204, 116 Stat. 745. The Public Company Accounting Reform and Investor Protection Act.
Ogden, R., Linacre, A., 2015. Wildlife forensic science: a review of genetic geographic origin assignment. Forensic Science International: Genetics 18, 152–159.
Oliveira, D. D., Baião, F. A., & Mattoso, M. (2010). Towards a taxonomy for cloud computing from an e-science perspective. In Cloud computing (pp. 47-62). Springer, London.
Olympic Ins. Co. v. H. D. Harrison, (1969). Inc., 418 F.2d 669, 670 (5th Cir. 1969).
Omolara, A.E., Jantan, A., Abiodun, O.I., Singh, M.M., Anbar, M., Kemi, D.V., 2018. State-of-the-art in big data application techniques to financial crime: a survey. International Journal of Computer Science and Network Security 18 (7), 6–16.
Omolara, A.E., Alabdulatif, A., Abiodun, O.I., Alawida, M., Alabdulatif, A., Arshad, H., 2022. The internet of things security: A survey encompassing unexplored areas and new insights. Computers & Security 112, 102494.
Overill, R.E., 2009. Development of masters modules in computer forensics and cybercrime for computer science and forensic science students. International Journal of Electronic Security and Digital Forensics 2 (2), 132–140.
Parker, D., 2005. Windows ntfs alternate data streams. Security. Focus 16.
Pasquier, T., Han, X., Goldstein, M., Moyer, T., Eyers, D., Seltzer, M., & Bacon, J. (2017, September). Practical whole-system provenance capture. In Proceedings of the 2017 Symposium on Cloud Computing (pp. 405-418).
Pasquier, T., Han, X., Moyer, T., Bates, A., Hermant, O., Eyers, D., ... & Seltzer, M. (2018, October). Runtime analysis of whole-system provenance. In Proceedings of the 2018 ACM SIGSAC conference on computer and communications security (pp. 1601-1616).
Patel, S., Rahevar, M., & Parmar, M. (2020). Data Provenance and Data Lineage in the Cloud: A Survey.
Perera, C., Ranjan, R., Wang, L., Khan, S.U., Zomaya, A.Y., 2015. Big data privacy in the internet of things era. IT Professional 17 (3), 32–39.
Pichan, A., Lazarescu, M., Soh, S.T., 2015. Cloud forensics: Technical challenges, solutions and comparative analysis. Digital investigation 13, 38–57.
Pignotti, E., Corsar, D., Edwards, P., 2011. Provenance principles for open data. Nottingham, UK.
Pointcheval, D., Stern, J., 2000. Security arguments for digital signatures and blind signatures. Journal of cryptology 13 (3), 361–396.
Polyzos, G.C., Fotiou, N., 2017. Blockchain-assisted information distribution for the Internet of Things. In: In 2017 IEEE International Conference on Information Reuse and Integration (IRI). IEEE, pp. 75–78.
Pope, A.L., 1998. The CORBA reference guide: understanding the common object request broker architecture. Addison-Wesley Longman Publishing Co., Inc..
Potthoff, J., Tremouilhac, P., Hodapp, P., Neumair, B., Bräse, S., Jung, N., 2019. Procedures for systematic capture and management of analytical data in academia. Analytica chimica acta: X 1, 100007.
Prasad, K. M., Reddy, A., & Karthik, M. G. (2012). Flooding attacks to internet threat monitors (ITM): modeling and counter measures using botnet and honeypots. arXiv preprint arXiv:1201.2481.
Prayudi, Y., Sn, A., 2015. Digital chain of custody: State of the art. International Journal of Computer Applications 114 (5).
Quick, D., Choo, K.K.R., 2013. Forensic collection of cloud storage data: Does the act of collection result in changes to the data or its metadata? Digital Investigation 10 (3), 266–277.
Raghavendra, S., Srividya, P., Mohseni, M., Bhaskar, S. C., Chaudhury, S., Sankaran, K. S., & Singh, B. K. (2022). Critical Retrospection of Security Implication in Cloud Computing and Its Forensic Applications. Security and Communication Networks, 2022.
Rajbhandari, S., & Walker, D. W. (2004). Support for provenance in a service-based computing grid. In UK e-Science All Hands Meeting.
Ram, S., Liu, J., 2006. In: November). Understanding the semantics of data provenance to support active conceptual modeling. Springer, Berlin, Heidelberg, pp. 17–29.
Rani, A., Goyal, N., Gadia, S.K., 2022. Social data provenance framework based on zero-information loss graph database. Social Network Analysis and Mining 12 (1), 1–25.
Rasool, A., Jalil, Z., 2020. A review of web browser forensic analysis tools and techniques. Researchpedia Journal of Computing 1 (1), 15–21.
Razaque, A., Frej, M.B.H., Alotaibi, B., Alotaibi, M., 2021. Privacy Preservation Models for Third-Party Auditor over Cloud Computing: A Survey. Electronics 10 (21), 2721.
Reilly, C. F., & Naughton, J. F. (2009, February). Transparently Gathering Provenance with Provenance Aware Condor. In Workshop on the Theory and Practice of Provenance.
Reynolds, P., Killian, C. E., Wiener, J. L., Mogul, J. C., Shah, M. A., & Vahdat, A. (2006, May). Pip: Detecting the Unexpected in Distributed Systems. In NSDI (Vol. 6, pp. 9-9).
Rivest, R.L., Shamir, A., Adleman, L., 1978. A method for obtaining digital signatures and public-key cryptosystems. Communications of the ACM 21 (2), 120–126.

Rogers, M.K., Seigfried, K., 2004. The future of computer forensics: a needs analysis survey. Computers & Security 23 (1), 12–16.
Rosenthal, A., Seligman, L., Chapman, A., Blaustein, B., 2009. Scalable access controls for lineage. MITRE CORP BEDFORD MA.
Ruan, K., Carthy, J., Kechadi, T., Crosbie, M., 2011. Cloud forensics. In: IFIP International Conference on Digital Forensics. Springer, Berlin, Heidelberg, pp. 35–46.
Ruan, P., Chen, G., Dinh, T.T.A., Lin, Q., Ooi, B.C., Zhang, M., 2019. Fine-grained, secure and efficient data provenance on blockchain systems. Proceedings of the VLDB Endowment 12 (9), 975–988.
Ruan, P., Dinh, T.T.A., Lin, Q., Zhang, M., Chen, G., Ooi, B.C., 2021. LineageChain: a fine-grained, secure and efficient data provenance system for blockchains. The VLDB Journal 30 (1), 3–24.
Rupprecht, L., Davis, J.C., Arnold, C., Gur, Y., Bhagwat, D., 2020. Improving reproducibility of data science pipelines through transparent provenance capture. Proceedings of the VLDB Endowment 13 (12), 3354–3368.
Russell, D., Russell, D., Gangemi, G. T., & Gangemi Sr, G. T. (1991). Computer security basics. " O'Reilly Media, Inc.".
Ruth, P., Xu, D., Bhargava, B., Regnier, F., 2004. E-notebook middleware for accountability and reputation based trust in distributed data sharing communities. In: International Conference on Trust Management. Springer, Berlin, Heidelberg, pp. 161–175.
Sahu, S., Dhote, Y., 2016. A study on big data: Issues, challenges and applications. International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE) 4 (6), 10611–10616.
Sakka, M.A., Defude, B., Tellez, J., 2010. Document provenance in the cloud: constraints and challenges. In: Meeting of the european network of universities and companies in information and communication engineering. Springer, Berlin, Heidelberg, pp. 107–117.
Sakka, M.A., Defude, B., Tellez, J., 2012. A semantic framework for the management of enriched provenance logs. In: In: Advanced Information Networking and Applications (AINA), 2012 IEEE 26th International Conference on. IEEE, pp. 352–359.
Saleem, S. (2015). Protecting the integrity of digital evidence and basic human rights during the process of digital forensics (Doctoral dissertation, Department of Computer and Systems Sciences, Stockholm University).
Santra, P., Roy, A., Majumder, K., 2018. A Comparative analysis of cloud forensic techniques in IaaS. In: Advances in Computer and Computational Sciences. Springer, Singapore, pp. 207–215.
Savage, S., Wetherall, D., Karlin, A., Anderson, T., 2001. Network support for IP traceback. IEEE/ACM transactions on networking 9 (3), 226–237.
Schröder, D., Unruh, D., 2012. Security of blind signatures revisited. In: International Workshop on Public Key Cryptography. Springer, Berlin, Heidelberg, pp. 662–679.
Seltzer, M. I., Muniswamy-Reddy, K. K., Holland, D. A., Braun, U., & Ledlie, J. (2005 Jul). Provenance-aware storage systems. In: USENIX ATC, vol. 6.
Sheikh, U., Khan, A., Ahmed, B., Waheed, A., Hameed, A., 2018. Provenance inference techniques: Taxonomy, comparative analysis and design challenges. Journal of Network and Computer Applications 110, 11–26.
Sheng, Z., Yang, S., Yu, Y., Vasilakos, A.V., McCann, J.A., Leung, K.K., 2013. A survey on the ietf protocol suite for the internet of things: Standards, challenges, and opportunities. IEEE wireless communications 20 (6), 91–98.
Shetty, S., Red, V., Kamhoua, C., Kwiat, K., Njilla, L., 2017. Data provenance assurance in the cloud using blockchain. Disruptive Technologies in Sensors and Sensor Systems, 10206. SPIE, pp. 125–135.
Shibli, M.A., Masood, R., Habiba, U., Kanwal, A., Ghazi, Y., Mumtaz, R., 2014. Access control as a service in cloud: challenges, impact and strategies. In: Continued Rise of the Cloud. Springer, London, pp. 55–99.
Shvaiko, P., Euzenat, J., 2013. Ontology matching: state of the art and future challenges. Knowl. Data Eng. IEEE Trans. 25 (1), 158–176.
Sigwart, M., Borkowski, M., Peise, M., Schulte, S., & Tai, S. (2019, October). Blockchain-based data provenance for the Internet of Things. In Proceedings of the 9th International Conference on the Internet of Things (pp. 1-8).
Sikos, L.F., Philp, D., 2020. Provenance-aware knowledge representation: A survey of data models and contextualized knowledge graphs. Data Science and Engineering 5 (3), 293–316.
Silva, V., De Oliveira, D., Valduriez, P., Mattoso, M., 2016. Analyzing related raw data files through dataflows. Concurrency and Computation: Practice and Experience 28 (8), 2528–2545.
Simmhan, Y.L., Plale, B., Gannon, D., 2005. A survey of data provenance in e-science. ACM Sigmod Record 34 (3), 31–36.
Simmhan, Y.L., Plale, B., Gannon, D., 2005b. A survey of data provenance techniques. Computer Science Department, Indiana University, Bloomington IN 47405, 69.
Sion, R., & Winslett, M. (2006). Towards Regulatory Compliance in Data Management.Sion, R., & Winslett, M. (2006). Towards Regulatory Compliance in Data Management.
Sohal, M., Sharma, S., 2022. BDNA-A DNA inspired symmetric key cryptographic technique to secure cloud computing. Journal of King Saud University-Computer and Information Sciences 34 (1), 1417–1425.
Somani, G., Agarwal, A., Ladha, S., 2012. Overhead analysis of security primitives in cloud. In: In 2012 International Symposium on Cloud and Services Computing. IEEE, pp. 129–135.
Srinivasamurthy, S., & Liu, D. Q. (2010, November). Survey on cloud computing security. In Proc. Conf. on Cloud Computing, CloudCom (Vol. 10).
Stoykova, R., Andersen, S., Franke, K., Axelsson, S., 2022. Reliability assessment of digital forensic investigations in the Norwegian police. Forensic Science International: Digital Investigation 40, 301351.
Subramanya, S.R., Lakshminarasimhan, N., 2001. Computer viruses. IEEE potentials 20 (4), 16–19.
Suen, C.H., Ko, R.K., Tan, Y.S., Jagadpramana, P., Lee, B.S., 2013. S2logger: End-to-end data tracking mechanism for cloud data provenance. In: In 2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications. IEEE, pp. 594–602.
Suhail, S., Hong, C.S., Ahmad, Z.U., Zafar, F., Khan, A., 2016. Introducing secure provenance in iot: Requirements and challenges. In: In 2016 International Workshop on Secure Internet of Things (SIoT). IEEE, pp. 39–46.
Sultana, S., Bertino, E., Shehab, M., 2011. A provenance-based mechanism to identify malicious packet dropping adversaries in sensor networks. In: In 2011 31st International Conference on Distributed Computing Systems Workshops. IEEE, pp. 332–338.
Swanson, G.E., 1960. The birth of the gods: The origin of primitive beliefs, Vol. 93. University of Michigan Press.
Tan, W.C., 2007. Provenance in databases: Past, current, and future. IEEE Data Eng. Bull. 30 (4), 3–12.
Tan, Y.S., Ko, R.K., Holmes, G., 2013. Security and data accountability in distributed systems: A provenance survey. In: In 2013 IEEE 10th International Conference on High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing. IEEE, pp. 1571–1578.
Tavares, B., Correia, F.F., Restivo, A., Faria, J.P., Aguiar, A., 2018. A survey of blockchain frameworks and applications. In: International Conference on Soft Computing and Pattern Recognition. Springer, Cham, pp. 308–317.
Townend, P., Groth, P., Xu, J., 2005. A provenance-aware weighted fault tolerance scheme for service-based applications. In: Eighth IEEE International Symposium on Object-Oriented Real-Time Distributed Computing (ISORC'05). IEEE, pp. 258–266.
Toyoda, K., Mathiopoulos, P.T., Sasase, I., Ohtsuki, T., 2017. A novel blockchain-based product ownership management system (POMS) for anti-counterfeits in the post supply chain. IEEE access 5, 17465–17477.
Trenwith, P.M., Venter, H.S., 2014. A digital forensic model for providing better data provenance in the cloud. In: In 2014 Information Security for South Africa. IEEE, pp. 1–6.
Triantafyllou, A., Mattielli, N., Clerbois, S., Da Silva, A.C., Kaskes, P., Claeys, P., Brkojewitsch, G., 2021. Optimizing multiple non-invasive techniques (PXRF, pMS, IA) to characterize coarse-grained igneous rocks used as building stones. Journal of Archaeological Science 129, 105376.
Tribis, Y., El Bouchti, A., Bouayad, H., 2018. Supply chain management based on blockchain: A systematic mapping study. MATEC Web of Conferences, 200. EDP Sciences, p. 00020.
US-CERT, A. H. (2011). James Cebul a,"The Basics of Cloud Computing," US-C ERT.
Van Buskirk, E., Liu, V.T., 2006. Digital evidence: Challenging the presumption of reliability. Journal of Digital Forensic Practice 1 (1), 19–26.
Ventures, C. (2019). 2019 official annual cybercrime report.
Vicknair, C., Macias, M., Zhao, Z., Nan, X., Chen, Y., & Wilkins, D. (2010, April). A comparison of a graph database and a relational database: a data provenance perspective. In Proceedings of the 48th annual Southeast regional conference (pp. 1-6).
Wang, J., Crawl, D., Purawat, S., Nguyen, M., Altintas, I., 2015. Big data provenance: Challenges, state of the art and opportunities. In: In 2015 IEEE International Conference on Big Data (Big Data). IEEE, pp. 2509–2516.
Wang, C., Carzaniga, A., Evans, D., & Wolf, A. L. (2002, January). Security issues and requirements for internet-scale publish-subscribe systems. In Proceedings of the 35th Annual Hawaii International Conference on System Sciences (pp. 3940-3947). IEEE.
Wang, M., Zhang, Q., 2020. Optimized data storage algorithm of IoT based on cloud computing in distributed system. Computer Communications 157, 124–131.
Werder, K., Ramesh, B., Zhang, R., 2022. Establishing Data Provenance for Responsible Artificial Intelligence Systems. ACM Transactions on Management Information Systems (TMIS) 13 (2), 1–23.
Wong, S.C., Miles, S., Fang, W., Groth, P., Moreau, L., 2005. Provenance-based validation of e-science experiments. In: International Semantic Web Conference. Springer, Berlin, Heidelberg, pp. 801–815.
Woodman, S., Hiden, H., Watson, P., 2017. Applications of provenance in performance prediction and data storage optimisation. Future Generation Computer Systems 75, 299–309.
Woods, C. M. (2015). Implementing cyber coercion. NAVAL POSTGRADUATE SCHOOL MONTEREY CA.
Wylot, M., Cudré-Mauroux, P., Hauswirth, M., Groth, P., 2017. Storing, tracking, and querying provenance in linked data. IEEE Transactions on Knowledge and Data Engineering 29 (8), 1751–1764.
Xia, Q.I., Sifah, E.B., Asamoah, K.O., Gao, J., Du, X., Guizani, M., 2017. MeDShare: Trust-less medical data sharing among cloud service providers via blockchain. IEEE access 5, 14757–14767.
Xie, Y., Muniswamy-Reddy, K.K., Feng, D., Li, Y., Long, D.D., 2013. Evaluation of a hybrid approach for efficient provenance storage. ACM Transactions on Storage (TOS) 9 (4), 1–29.
Xu, S., Rogers, T., Fairweather, E., Glenn, A., Curran, J., Curcin, V., 2018. Application of data provenance in healthcare analytics software: information visualisation of user activities. AMIA Summits on Translational Science Proceedings 2018, 263.

Xue, C.T.S., Xin, F.T.W., 2016. Benefits and challenges of the adoption of cloud computing in business. International Journal on Cloud Computing: Services and Architecture 6 (6), 01–15.

Yan, Z., Zhang, P., Vasilakos, A.V. (2014). A survey on trust management for internet of things. J. Netw. Comput. Appl. 42, 120 –134 (2014).

Yang, J., Wen, J., Jiang, B., Wang, H., 2020. Blockchain-based sharing and tamper-proof framework of big data networking. IEEE Network 34 (4), 62–67.

Yaniv, Z., Lowekamp, B.C., Johnson, H.J., Beare, R., 2018. SimpleITK image-analysis notebooks: a collaborative environment for education and reproducible research. Journal of digital imaging 31 (3), 290–303.

Yao, Y., Gehrke, J., 2002. The cougar approach to in-network query processing in sensor networks. ACM Sigmod record 31 (3), 9–18.

Zafar, F., Khan, A., Suhail, S., Ahmed, I., Hameed, K., Khan, H.M., Anjum, A., 2017. Trustworthy data: A survey, taxonomy and future trends of secure provenance schemes. Journal of network and computer applications 94, 50–68.

Zargar, S.T., Joshi, J., Tipper, D., 2013. A survey of defense mechanisms against distributed denial of service (DDoS) flooding attacks. IEEE communications surveys & tutorials 15 (4), 2046–2069.

Zawoad, S., & Hasan, R. (2013). Cloud forensics: a meta-study of challenges, approaches, and open problems. arXiv preprint arXiv:1302.6312.

Zhang, K., 2013. Illegal Invasion of Computer Information Systems. In: Informatics and Management Science V. Springer, London, pp. 175–182.

Zhang, C., Chen, Y., 2020. A review of research relevant to the emerging industry trends: Industry 4.0, IoT, blockchain, and business analytics. Journal of Industrial Integration and Management 5 (01), 165–180.

Zhang, Q., Cheng, L., Boutaba, R., 2010. Cloud computing: state-of-the-art and research challenges. Journal of internet services and applications 1 (1), 7–18.

Zhang, L., Li, J., Li, P., Lu, X., Gong, M., Shen, P., Schuller, B.W., 2022. MEDAS: an open-source platform as a service to help break the walls between medicine and informatics. Neural Computing and Applications 34 (8), 6547–6567.

Zhou, W., Cronin, E., & Loo, B. T. (2007). Provenance-aware declarative secure networks.

Zhou, W., Sherr, M., Tao, T., Li, X., Loo, B. T., & Mao, Y. (2010, June). Efficient querying and maintenance of network provenance at internet-scale. In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data (pp. 615-626).

Zhou, W., Fei, Q., Narayan, A., Haeberlen, A., Loo, B. T., & Sherr, M. (2011, October). Secure network provenance. In Proceedings of the twenty-third ACM symposium on operating systems principles (pp. 295-310).