Big Data Issues:
Analytics and Security

B

Dina Darwish
Ahram Canadian University, Egypt

INTRODUCTION

Big data is a collection of structured, semi-structured, and unstructured data that is gathered by businesses and may be mined to be used in advanced analytics applications using machine learning and predictive modelling. Big data is used by businesses to enhance operations, deliver better customer service, develop individualized marketing campaigns, and carry out other tasks that can ultimately boost sales and profits. Because they can act more quickly and with greater knowledge, businesses who use it efficiently may have a competitive advantage over those that don't. Data scientists and other data analysts need to have a thorough comprehension of the available data and a clear understanding of what they're looking for in it in order to provide reliable and pertinent results from big data analytics applications. As a result, a vital preparatory stage in the analytics process is data preparation, which involves profiling, cleansing, validation, and transformation of data sets. Securing Big Data, which frequently contains sensitive information, is another crucial concern that needs to be addressed.

This chapter is going to address two important issues for Big Data; which are analytics and security. Since, today companies and data scientists work on Big Data analytics to retrieve relevant information, and provide essential statistics to help companies and research labs in finding solutions to some questions, and improve their sales and competitiveness in the market, and at the same time, keeping these information secured. The objectives of this article are:

1. Concentrating on Big Data security related issues
2. Presenting various solutions to Big Data security challenges
3. Clarifying importance of Big Data analytics, especially for companies
4. Focusing on several solutions to issues related to Big Data security
5. Providing strategies in general for managing effectively Big Data
6. Discussing conclusion, and future research directions.

This chapter is organized as follows; the background represents the second section, where literature review is mentioned, the third section is composed of focus of the chapter, which is definition of Big Data, its importance, applications, and analytics, section 4 focuses on solutions and recommendations for Big Data security and analytics challenges, and finally, comes conclusion and future research directions.

BACKGROUND

Research needs to be done on two crucial Big Data issues: analytics and security. Some research on these two issues are mentioned in the following sections. Social networks, such as Facebook and Twit-

DOI: 10.4018/978-1-6684-7366-5.ch020

ter are major producers of Big Data. Social networks like Facebook and Twitter are leading producers of big data. Social Network topology has strong impact on physical technological networks as most of the traffic is contributed by these social network sites and related ones. Some approaches and examples of the use of social network analysis in the design of technology networks and vice versa are explored in research (Cheng et al., 2013). A user can be a member of multiple networks at the same time, and these networks can combine to produce a composite social network in which the person's activity varies across networks. Moreover, the user may share similar latent interests with other users throughout these networks. E. Zhong (Zhong et al.,2012) proposed a model for adaptive transfer of knowledge from composite social networks to forecast human behavior for use in social marketing, service suggestions, and personalization. Using personal ad hoc clouds of users in social networks to address big data processing difficulties by leveraging the social network paradigm for generating information from big data was discussed in research (Tan et al., 2013). The combination of IoT, big data analytics, and complicated event processing techniques is suggested by authors (Tasweef et al., 2015) as a solution to the main problems with handling data in the healthcare industry. They proposed an all-encompassing healthcare system that could carry out tasks like drug detection, monitor patients from a distance, help with health insurance settlement, and advance the therapeutic outcomes.

Through the use of big data analytics, organizations can obtain useful data and patterns that could have an impact on their operations (Gandomi & Haider, 2015). Therefore, in order to determine the relationships between features and predict future observations, extensive data analysis is required. Big data analytics refers to methods used to draw conclusions from enormous databases (Labrinidis & Jagadish, 2012). The outcomes of big data analytics can enhance decision-making and boost organizational effectiveness. In order to extract knowledge from the data, many analytical techniques are developed, such as; Descriptive analytics is focused with examining historical data of a business to describe what happened in the past; (Joseph & Johnson, 2013), To forecast potential outcomes, predictive analytics focuses on various statistical modelling and machine learning techniques (Waller & Fawcett, 2013). Descriptive and predictive analytics are combined in prescriptive analytics to suggest the best action for enhancing company procedures (Joseph & Johnson, 2013).

Vaishya and his colleagues (Vaishya et al., 2020) explored the primary uses of AI for preventing and combating Coronavirus Disease through the use of big data analytics (COVID-19). The scientists identified seven uses of AI for the COVID-19 pandemic, including: disease detection, patient treatment monitoring, contact tracing, cases and deaths prediction, medicine manufacture, workload reduction, and disease prevention. The following, however, are factors that this paper neglects to consider: the small number of papers examined; the lack of a statement regarding the study selection procedure; and the absence of any qualitative characteristics. Furthermore, a thorough taxonomy based on AI methods was not provided. Finally, Pham and his colleagues (Pham et al., 2020) described how to organize and evaluate the enormous volume of data obtained by the COVID-19 disease using big data and AI techniques. Five categories—including COVID-19 outbreak prediction, viral tracking, diagnosis and treatment, and drug discovery—are taken into consideration when evaluating certain big data tools. The associated problems with the solutions under examination were then highlighted.

Using machine learning techniques, Kaur and his colleagues (Kaur et al., 2018) suggested a unique approach for intelligent healthcare information systems. There are four layers in the suggested model. Diverse data sources are handled by the data source layer. The storage optimization procedure is controlled by the data storage layer. To make the best use of system resources, a number of techniques have been applied, including normalization and indexing. The data security layer uses a variety of data security and privacy techniques, including data masking, granular control over data access, activity monitor-

ing, dynamic encryption, and endpoint validation. Last but not least, machine learning techniques were utilized at the application layer for early disease diagnosis Based on the trial results of the article, fuzzy logic and information theory increased the proposed model's accuracy. Nair and his colleagues (Nair et al., 2018) developed a novel health status prediction system using big data streams and machine learning models. The demonstrated solution was created with Apache Spark and installed in the cloud.

Utilizing historical data, predictive analytics makes predictions about consumer trends and behavior (Mosavi & Vaezipour, 2013). It is the process of forecasting future trends using historical data. To find patterns and learn from past data, this research uses statistical models and machine learning techniques (Shmueli & Koppius, 2011). The method of using machine learning to examine data and create predictions is known as predictive analysis (Puri, 2013). Sixty seven percent of firms want to use predictive analytics to develop more strategic marketing campaigns in the future, and sixty eight percent see the competitive advantage is the main advantage of using predictive analysis (Millard, 2013). Predictive search, price management, and product recommendation are three general uses of predictive analysis in e-commerce. A huge e-commerce site typically sells thousands of goods and services. For customers, having to navigate and browse through thousands of products on a website could be very frustrating. An E-Commerce site or application may now instantly discover or forecast products that closely match the consumer's preferences thanks to the introduction of recommender systems (Sarwar et al., 2002).

Additionally, security and privacy concerns related to big data are becoming increasingly urgent (Agrawal et al., 2011). As an example, people can now easily share and distribute valuable copyrighted digital contents thanks to social media. As a result, copyright infringing activities such illegal copying, harmful distribution, unauthorized access and usage, and unrestricted sharing of digital content will increase significantly in frequency. Big Data should include dependable solutions to support author privacy and copyrights in order to lessen these issues (Marques & Serrão, 2013).

Additionally, users lose control over their data and content, seriously compromising their own privacy, as they share more and more personal information and user-generated material on social networks and through cloud services using their mobile devices and laptops. A potentially effective strategy is to increase the level of uncertainty for attackers by dynamically altering system features in a concept known as a "cyber moving target" (MT) (Okhravi et al., 2014). This strategy gives an overview of several MT strategies, weigh their advantages and disadvantages. To make the analytics process real-time and so increase the value of the analysis, it is necessary to identify the life cycle of the data, the value it may bring, and the computing process (Chen et al., 2014). Big data is not always better, so appropriate data filtering techniques can be created to guarantee the accuracy of the data (Boyd & Crawford, 2012). The accessibility of accurate and reliable data is a significant problem as well. The majority of the time, sparse data with unclear distributions lead to false conclusions.

To process huge amounts of data efficiently, big data needs new kinds of technologies. Data fusion and integration, genetic algorithms, machine learning, signal processing, simulation, natural language processing, time series analytics, and visualization are some of the technologies used (Chen & Zhang, 2014; Kaisler et al., 2013; Song et al., 2013). Koscielniak and Puto (Kościelniak & Puto, 2015) pointed out the growing significance of how information is arranged and used for decision-making. The findings of a poll carried out by McKinsey in 2011 (McKinsey & Company, 2011), which indicate that the analysis of Big Data may eventually become a crucial foundation for competition, productivity, and innovation, further corroborate this viewpoint. Businesses can gain a competitive edge by utilizing the outcomes of big data analytics (Kościelniak & Puto, 2015). However, leveraging conventional data to enhance decision-making is not new. For instance, Ziora (Ziora, 2015) describes how to determine users' favorite product features based on data collected during the product evaluation phase (Barbacioru,

2014). For large data analytics, a number of solutions have been put forth, which can be broken down (Pospiech & Felden, 2012) into three categories:

(1) processing/computing (Apache Hadoop, 2015), Nvidia CUDA (Cuda, 2015), or Twitter Storm (Apache Storm, 2015),
(2) storage (Titan or HDFS), and
(3) analytics (MLPACK (Curtin et al., 2013) or Mahout (Apache Mahout, 2015)). Even though there are commercial programs for data analysis, most studies on traditional data analysis concentrate on designing and developing effective "means" to extract usable information from the data.

MAIN FOCUS OF THE CHAPTER

Definition of Big Data

Big data is a collection of structured, semi-structured, and unstructured data that is gathered by businesses and may be mined for use in advanced analytics applications like machine learning and predictive modelling.

Big data can now refer to vast data sets, systems, and solutions created to manage such enormous data assemblages, as well as the area of computers committed to its development (Merriam Webster dictionary, 2023). The term "Big Data" refers to the vast amount of data in today's networked, digital, sensor-rich, information-driven world. Data analytics scientific and technological advancements are being outpaced by data expansion (NIST Framework, 2019). Big data is the term used to describe gathered data sets that are so huge and complicated that they need to be processed by modern technologies like artificial intelligence. The information is gathered from numerous sources (European parliament, 2021).

Big data are information collections that are too big or complex to manage, analyse, or utilise in the usual ways (Oxford dictionary, 2023).

The term "Big Data" describes the phenomena of the exponential expansion of business data in the twenty-first century as well as the problems it poses, such as the comprehensive gathering, storage, administration, and analysis of all the data that a company utilises or owns (Informatica, 2023). The enormous amounts of data that have been gathered over time and are challenging to handle and evaluate using standard database administration techniques. The data are analyzed for marketing trends in business, industry, medicine, and science. Business transactions, emails, pictures, surveillance footage, browsing patterns, activity logs, unstructured text from blogs and social media, as well as the enormous volumes of data that may be gathered from sensors of all kinds, are among the data categories (PC magazine, 2023).

Large datasets with high volume, variety, velocity, and/or variability make up big data, which necessitates a scalable architecture for effective manipulation, analysis, and storage (De Mauro et al., 2015).

We can confirm that the core of the concept of big data can be expressed by the following: volume, velocity, and variety, to describe the characteristics of the information involved; specific technology and analytical methods, to clarify the specific requirements strictly needed to make use of such Information; as well as transformation into insights and subsequent creation of economic 'value' (NIST Framework, 2019).

Together with tools that support big data analytics uses, systems that process and store big data have become a common part of data management architectures in businesses. The three V's are frequently used to describe big data:

- the volume of data in various forms;
- the enormous variety of data types that big data systems often hold;
- the velocity with which a lot of data is created, gathered, and processed.

Doug Laney, a consultant at Meta Group Inc., first defined these traits in 2001; Gartner further popularized them after acquiring Meta Group in 2005. Veracity, value, and variability are a few additional V's that have been added to various definitions of big data more recently. Big data deployments can involve terabytes, petabytes, and even exabytes of data that have been created and gathered over time, even though big data does not have any exact volume of data.

Importance of Big Data

Big data is used by businesses to enhance operations, deliver better customer service, develop individualized marketing campaigns, and carry out other tasks that can ultimately boost sales and profits. Because they can act more quickly and with greater knowledge, businesses who use it efficiently may have a competitive advantage over those that don't use it. Big data, for instance, offers insightful information about customers that businesses can utilize to improve their marketing, advertising, and promotions and boost customer engagement and conversion rates. Businesses can become more responsive to customer demands and needs by analyzing historical and real-time data to gauge the changing preferences of consumers or corporate buyers. Big data is also utilized by doctors to assist in the diagnosis of illnesses and medical problems in patients as well as by medical researchers to find disease indicators and risk factors. Additionally, healthcare institutions and governmental organizations receive up-to-date information about infectious disease threats or outbreaks via a combination of data from electronic health records, social media platforms, the web, and other sources. Here are some further instances of how businesses use big data:

- Big data is utilized by utilities to monitor electrical grids and by oil and gas corporations to locate possible drilling sites and follow pipeline activity in the energy sector.
- Big data platforms are used by financial services companies for risk management and in-the-moment market data analysis.
- Big data is used by manufacturers and transportation firms to manage their supply networks and improve delivery routes.
- Emergency response, crime prevention, and smart city programs are further government uses.

Examples of Big Data

Big data is derived from a variety of sources, including customer databases, transaction processing systems, documents, emails, medical records, clickstream logs on the internet, mobile apps, and social networks. It also includes data that is produced by machines, like network and server log files, as well as data from sensors on industrial machinery, internet of things devices, and manufacturing machines. Big data environments frequently include external data on consumers, financial markets, weather and traffic conditions, geographic information, scientific research, and more in addition to data from internal systems. Big data applications frequently use streaming data that is processed and gathered continuously, including images, videos, and audio files.

Big Data analytics

Data scientists and other data analysts need to have a thorough comprehension of the available data and a clear understanding of what they're looking for in it in order to provide reliable and pertinent results from big data analytics applications. As a result, a vital first stage in the analytics process is data preparation, which involves profiling, cleansing, validation, and transformation of data sets. Using technologies that offer big data analytics features and capabilities, multiple data science and advanced analytics disciplines can be deployed to run different applications once the data has been obtained and readied for analysis. These fields include text mining, predictive modelling, data mining, statistical analysis, streaming analytics, and machine learning, including its deep learning branch.

The various areas of analytics that can be carried out using big data sets include the following, using customer data as an example:

- **comparative examination.** To compare a company's goods, services, and branding to those of its rivals, this looks at customer behavior metrics and real-time customer engagement.
- **watching social media.** This examines what customers are saying on social media about a company or product in order to find potential issues and pinpoint the right target market.
- **analytics in marketing.** This delivers data that can be utilized to enhance advertising campaigns and sales promotions for goods, services, and company endeavors.
- **Sentimental evaluation.** Customer satisfaction levels, attitudes toward a company or brand, potential problems, and ways to enhance customer service can all be discovered through analysis of the data gathered on customers.

Figure 1. Areas of Big Data analytics using customer data

Figure 1 illustrates areas of Big Data analytics using customer data.

B

Big Data Analytics Case Studies

Starbucks: The coffee giant is at the forefront of leveraging big data and artificial intelligence to assist guide marketing, sales, and business choices with 90 million transactions each week across 25,000 shops globally. Starbucks posseses the unique purchasing data of millions of customers through its well-known loyalty card program and mobile application. The business forecasts purchases and emails specific offers to customers based on their expected preferences using this data and business intelligence (BI) technologies. This system raises sales volumes and encourages returning consumers to visit its locations more regularly. The same data that enables Starbucks to propose new goods to try also enables the business to provide tailored offers and discounts that go far beyond a particular birthday deal. (Systems plus, 2022).

Companies for airline solutions: BI speeds up for business insights. For airlines, hotels, and other businesses in the travel sector, Airline Solutions offers booking tools, revenue management, web- and mobile itinerary tools, as well as other technology. Additionally, the clients of the airline solution provider required cutting-edge solutions that could offer real-time data on customer behaviour and actions. To store the massive amounts of data, they created an enterprise travel data warehouse (ETDW). (Systems plus, 2022).

Walmart's use of Big Data: Big data is being used by Walmart to analyse the abundant information. That is pouring through its processes. Big data makes it possible to see the workflow at its pharmacies, distribution hubs, and retail locations in real time. The shopping experience at Walmart is improved, optimised, and customised in the following ways:

- Improve the effectiveness of Walmart pharmacies.
- Control the supply chain.
- Make the buying experience more individualised.
- Help enhance checkout in stores.
- Maximise the variety of products. (Techvidvan, 2023).

eBay and Big Data: Located in San Jose, California, eBay is a transnational American e-commerce company. Currently, Hortonworks HDF, Apache Spark, and Kafka are being used by eBay. Additionally, Presto, a Hadoop-based interactive query engine, is utilised. Big data is used by the eBay website for a variety of purposes, including assessing site performance and uncovering fraud. In order to encourage customers to make additional purchases on the website, it also used big data to evaluate user data. On its website, eBay has about 180 million active buyers and sellers. (Techvidvan, 2023).

Amazon: American multinational technology business Amazon is headquartered in Seattle, USA. It began as an online bookshop, but it now focuses on regulation and collaboration, which has generated income and boosted customer satisfaction. Over more than 1,400,000 servers, it is home to an estimated 1,000,000,000 gigabytes of data. Amazon is constantly in front when it comes to understanding its customers because to its innovation in data science and big data. (Projectpro, 2023).

Education, AI, and Big Data: One of the world's top publishers of medical and scientific information is Elsevier. Elsevier has used AI and Big Data to enhance its operations throughout the years. Due to the enormous amount of data the organisation has gathered over the course of its 140-year history, advanced analytics tools utilising Big Data and AI have been developed. (Mygreatlearning, 2020).

N-iX company is well-respected in the worldwide tech industry; it has been named one of Clutch's top software development providers, included in IAOP's Global Outsourcing 100 for six consecutive years. Besides, the company has expertise in the most appropriate tech stack for putting Big Data engineering, BI, Data Science, and AI/Machine Learning solutions into practise. N-iX works with Fortune 500 companies to help them make the most of big data and predictive analytics in their businesses. N-iX is in partnership with Fortune 500 companies helping them benefit from big data and predictive analytics in supply chain management (N-IX, 2020).

Security in Big Data

Big data analytics use is currently accelerating across all industries. Big data offers great opportunities to successfully enable a business to go ahead. The problem, though, is that platforms for big data analytics are typically loaded with a massive amount of products, partners, customers, and other data. This data typically lacks adequate data protection, which presents a wonderful opportunity to attackers.

Definition of Big Data Security

Big data security can be referred to as the instruments and controls that are employed to protect both data and analytics operations. Big data security's primary objective is to offer defense against assaults, thefts, and other nefarious actions that could compromise priceless data. The companies that use the cloud confront a variety of big data security challenges. This difficult threat includes online data theft, malware, and distributed denial of service (DDoS) attacks that could bring down a system. These risks could have a significant financial impact on a company, including losses, legal fees, fines, and punishments. Big data privacy and security concerns provide a challenge that must be overcome. With the aid of the proposed security intelligence model, intelligent analytics have been applied to improve security.

The goal of big data security is straightforward: prevent unwanted access and incursions via firewalls, robust user authentication, end-user education, intrusion protection systems (IPS), and intrusion detection systems (IDS). Encrypt data while it is in transit and at rest to prevent access. This method for network security seems standard. But because security technologies must function during three different data stages that aren't all available in the network, big data settings introduce an additional layer of security. The first is data ingress, the second is stored data, and the third is data output, which refers to the information that is sent to apps and reports.

Stage 1: Data Ingress. Many different sources and data types produce big data. Customer relation management (CRM) or enterprise risk management (ERM) data, transactional and database data, and enormous amounts of unstructured data like email or social media posts are all examples of user-generated data. The vast world of machine-generated data can be accessed, which includes logs and sensors. This data must be protected as it is being transferred from the sources to the platform.

Stage 2: Stored Data. It takes sophisticated security toolkits, such as encryption at rest, strong user authentication, intrusion protection, and planning, to protect stored data. Additionally, it is needed to use a distributed cluster infrastructure with plenty of servers and nodes to operate security toolkits. Additionally, log files and analytics tools must be protected by security tools while they are being used inside the platform.

Stage 3: Data output. The ability to conduct insightful analytics across enormous data volumes and many types of data is the primary driver behind the big data platform's complexity and cost. Applications, reports, and dashboards can use the findings of these analytics. It is crucial to encrypt both ingress

and output since this highly valuable intelligence makes a rich target for intrusion. At this step, it is required to ensure compliance by ensuring that results being sent to end users do not contain regulated data. Data is routed through a convoluted path and, theoretically, could be vulnerable at more than one point, which is one of the challenges of Big Data security.

Responsibility of Big Data Security

A big data deployment spans several business divisions. The big data deployment is a team effort involving IT, database administrators, programmers, quality testers, InfoSec, compliance officers, and business units. Policies, procedures, and security software that successfully defend the big data deployment against malware and unauthorized user access are the responsibility of IT and InfoSec. To safeguard compliance, compliance officers must collaborate closely with this team. For instance, results sent to a quality control team should automatically be extracted without credit card numbers. To protect their databases, DBAs should collaborate closely with IT and InfoSec.

Lastly, end users have equal responsibility for safeguarding corporate data. Contrary to popular belief, big data platforms are just as susceptible to malware and intrusion as any other type of data, despite the fact that many businesses use them to identify anomalies associated with intrusion. A straightforward email is one of the simplest ways for hackers to access networks, including big data platforms. Despite the fact that the majority of users are aware to delete the standard embarrassing attempts from Nigerian princes and bogus FedEx shipments, some phishing assaults are really smart. Never undervaluing the power of an email, whether managing security for a big data platform or an end user looking through his inbox. Protect your big data platform from both high and low dangers, and it will benefit the company for many years.

Issues with Big Data security

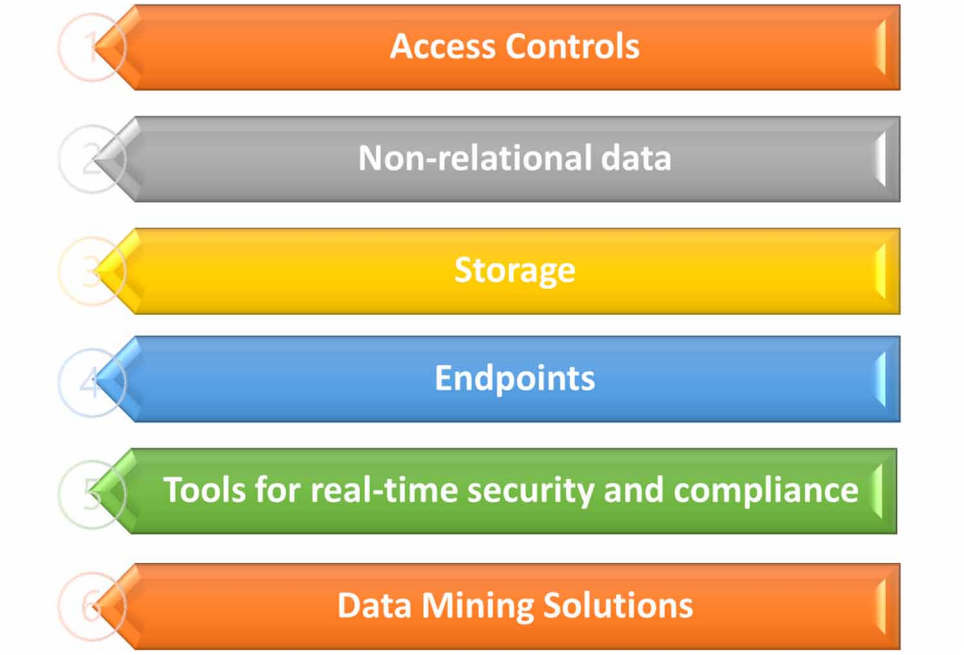
Issues related to securing Big Data can be summarized as follows:

- **Access Controls:** Having a system that is one hundred percent secure is crucial for a company. Only authenticated users should be given permission to exchange data. Access control must be set up so that it cannot be compromised by attackers, hackers, or other malicious actions. However, because it requires a significant financial outlay and ongoing maintenance, creating a fully secure and robust access control is a major challenge for enterprises.
- **Non-relational data stores:** Non-relational databases, such as NoSQL, typically don't come with security built in.
- **Storage:** Big data architecture uses multiple tiers to store data. The performance and cost of its storage are determined by business requirements. For instance, flash media is typically used to store high-importance data. Locking down storage therefore requires developing a tier-aware strategy.
- **Endpoints:** Security solutions that typically gather logs from endpoints must verify the veracity of those endpoints otherwise the analysis won't be very useful.
- **Tools for real-time security and compliance:** Real-time tools typically produce a lot of information. Finding a way to disregard inaccurate or incomplete information is crucial. So, that human resources can be concentrated on significant information or real breaches.

- **Data mining solutions:** Data mining remedies typically discover a pattern that denotes business tactics. For this reason, it must be protected from threats both inside and outside the organization.

Figure 2 illustrates Big Data security issues.

Figure 2. Big Data issues



Challenges Associated with Big Data Security

Users frequently face difficulties when designing a big data architecture due to problems with processing capacity. Big data systems must be customized to an organization's specific requirements; this is a do-it-yourself project that calls for IT and data management teams to assemble a specific set of technologies and tools. Compared to the abilities that database administrators and developers who work primarily with relational software typically possess, deploying and managing big data systems also calls for new abilities.

Using a managed cloud service can help with both of those problems, but IT managers must keep a close eye on cloud usage to ensure costs don't spiral out of control. Additionally, moving processing workloads and data sets from on-premises to the cloud is frequently a challenging process.

Making the data available to data scientists and analysts is another problem in managing big data systems, particularly in remote situations with a variety of different platforms and data repositories. Data management and analytics teams are increasingly creating data catalogues that include metadata management and data lineage functions to aid analysts in finding pertinent data. Big data integration is frequently a challenging process, especially when data variety and velocity are concerns.

There are a number of obstacles to big data security that could jeopardize it. Remember that these difficulties are not at all exclusive to on-premise big data platforms. They also apply to clouds. Take nothing for granted when hosting a big data platform in the cloud. Strong security service level agreements can help users and providers meet these same challenges. Common challenges in Securing Big Data are:

- Newer technologies under active development include advanced analytical tools for unstructured big data and non-relational databases (NoSQL). It can be challenging for security tools and procedures to safeguard these new toolkits. Data entry and storage are properly protected by sophisticated security measures. They might not, however, have a same effect on the data supplied from various analytics systems to various locations.
- Administrators of big data may choose to mine data without asking or being informed. Security solutions need to monitor and alert on questionable access regardless of the source, whether it's driven by curiosity or criminal profit.
- Terabytes to petabytes in size, a Big data installation is too huge for ordinary security audits. Due to the fact that the majority of big data platforms are cluster-based, numerous vulnerabilities are introduced across numerous servers and nodes.
- Big data owners run the risk of data loss and exposure if they don't routinely update the environment's security.
- Big data security experts need to continuously update their knowledge regarding cleanup and removal of malware and threats.

SOLUTIONS AND RECOMMENDATIONS

Big Data Collection Practices and Regulations

Big data misuse has the potential to grow along with its collection and use. The General Data Protection Regulation (GDPR), a data privacy law that went into effect in May 2018, was approved by the European Union in response to public outcry over data breaches and other violations of individual privacy. GDPR places restrictions on the kinds of personal data that organizations can collect and calls for either individual consent or adherence to other predetermined grounds. Additionally, it contains a right-to-be-forgotten clause that enables EU citizens to request that companies delete their data.

Although there are no comparable federal legislation in the United States, the California Consumer Privacy Act (CCPA) attempts to provide Californians more control over how their personal information is collected and used by businesses operating there. The CCPA was enacted in 2018 and went into effect on January 1, 2020. Businesses must carefully control the big data collection process to make sure they are in compliance with such legislation. To identify regulated data and keep unauthorized employees from accessing it, controls must be put in place.

Storage and Processing of Big Data

A data lake is frequently used to store Big Data. While data lakes can handle a variety of data types and are often based on Hadoop clusters, cloud object storage services, NoSQL databases, or other big data platforms, data warehouses are frequently built on relational databases and only hold structured data.

Many Big Data environments combine multiple systems in a distributed architecture; for example, a central data lake might be integrated with other platforms, including relational databases or a data warehouse. For specific analytics uses, the data in big data systems may be left in its unprocessed state and subsequently filtered and organized. In other instances, it is preprocessed using software for data preparation and mining to make it ready for applications that are run on a regular basis.

The underlying compute infrastructure faces significant demands as a result of big data processing. Clustered systems, which use tools like Hadoop and the Spark processing engine to distribute processing workloads across hundreds or thousands of commodity servers, frequently provide the necessary computing power. Getting that kind of processing capability in a cost-effective method is a challenge. As a result, big data systems frequently reside in the cloud. Organizations can use managed Big-data as a service from cloud providers or deploy their own cloud-based systems. Cloud users can scale up the required number of servers just long enough to accomplish big data analytics tasks. The firm just pays for the storage and compute time it uses, and the cloud instances may be shut off until they're needed again.

Big Data Management Technologies

Hadoop, an open source distributed processing platform launched in 2006, initially sat at the center of most Big Data infrastructures. The development of Spark and other processing engines pushed MapReduce, the engine integrated into Hadoop, farther to the side. The end result is an ecosystem of Big data technologies that can be used to various tasks but are frequently utilized in tandem. Many of those technologies are bundled together in big data platforms and managed services provided by IT vendors, primarily for usage in the cloud. At the moment, those contain the following products, listed alphabetically:

- Kindle EMR (formerly Elastic MapReduce)
- Google Cloud Platform
- Cloud Dataproc by Google
- Ezmeral Data Fabric by HPE (formerly MapR Data Platform)
- Google Cloud HDInsight

In addition to Hadoop and Spark, the following types of technologies are also available to businesses who want to install big data systems on their own, whether locally or in the cloud:

- cloud object storage systems like Amazon Simple Storage Service (S3), Google Cloud Storage, and Azure Blob Storage, as well as storage repositories like the Hadoop Distributed File System (HDFS);
- cluster management systems like Kubernetes, Mesos, and YARN, the built-in resource manager and job scheduler for Hadoop that goes by the name alone but actually stands for Yet Another Resource Negotiator;
- stream processing engines, including the Spark Streaming and Structured Streaming modules as well as Flink, Hudi, Kafka, Samza, and Storm;
- NoSQL databases that use diverse technologies, such as Cassandra, Couchbase, CouchDB, HBase, MarkLogic Data Hub, MongoDB, Neo4j, and Redis;
- platforms for data lakes and data warehouses, including Kylin, Snowflake, Amazon Redshift, Delta Lake, and Google BigQuery; and
- SQL query tools like Trino, Drill, Hive, Impala, and Presto.

Big Data analytics

The employees tasked with maintaining and interpreting the data will ultimately determine the commercial value and benefits of Big data efforts. Big Data can be contrasted with small data, a term that's

occasionally used to describe data sets that can be used for self-service Business Intelligence (BI) and analytics. Small data refers to data sets that can be easily used for predictive analytics applications or help businesses deploy a suitable infrastructure for big data projects. “Big Data is for machines; tiny data is for people,” goes a frequent maxim. Big Data deployments are important targets for would-be attackers, so big data security is a constant worry. A single ransomware assault may make your big data deployment exposed to ransom demands. Even worse, an unauthorized user may obtain access to big data to delete and sell vital information. Losses may be significant. This cause the risk of having IP sold to unauthorized parties worldwide, receiving regulatory penalties and judgements, and suffering significant reputational damage.

Big data platforms need to be secured using a combination of established security technologies, recently created toolkits, and sophisticated methods for security monitoring across the course of the platform’s lifetime.

Big Data Security Solutions

Big Data Security Trends

The explosion of Big Data that powers smart technology and the expanding drive for people to own and control how their personal data is used are two of the major developments in the field of Big data, but they are also somewhat at odds with one another. Terabytes of data containing very sensitive personal information being collected by technologies like IoT, artificial intelligence, machine learning, and even customer relationship management (CRM) databases. Big data in the form of personal information is helpful for businesses looking to better target their products and services at their target market, but it also means that all businesses and outside vendors are now accountable for the proper use and management of personal information. The majority of firms strive to adhere to consumer data rules and regulations as big data and its enterprise use cases develop, yet their security flaws make data vulnerable to breaches. Take a look at some of the most prominent big data trends, crucial security gaps that many businesses have, and advice for ensuring Big Data security:

- ***Upgrade the distributed and cloud security infrastructure***. Many businesses are switching to cloud and data fabric infrastructures that enable greater data storage scalability as a result of the growth of big data. Because cloud security features are frequently configured incorrectly and left vulnerable, traditional security principles are frequently used to construct cloud security. Consult cloud and storage providers to learn more about their offerings, find out if a security solution is built in, and find out if they or a third party partner recommend any additional security tools.
- ***Establish policies and practices for mobile device management.*** IoT and other mobile devices rank among the top sources and receivers of Big Data, but because so many of these gadgets are utilized in daily life, they also present a number of security flaws. Establish stringent guidelines for your employees’ use of company data on their personal devices, and be sure to implement extra security layers to control which devices have access to important information.
- ***Training And Best Practices For Data Security.*** Big data is frequently hacked as a result of an effective phishing attack or other targeted attack against an unaware employee. Set up many layers of authentication protection to restrict who may access sensitive data storage, and again, educate staff on typical socially engineered assaults and what they look like.

Big Data Security Technologies

These large data security tools are all not brand-new. Their scalability and capacity to protect numerous types of data at various stages are novel features. The Big Data security technologies can be:

- ***Encryption:*** To secure data across enormous data volumes, encryption techniques must be able to safeguard it both in transit and at rest. Additionally, encryption must function on a wide range of data types, including both user- and machine-generated data. Additionally, encryption tools must function on popular big data storage formats like relational database management systems (RDBMS), non-relational databases like NoSQL, and specialized filesystems like the Hadoop Distributed File System (HDFS) as well as with a variety of analytics toolkits and the output data they produce.
- ***Centralized key management:*** is a security best practice that has been around for many years. Big data settings, particularly ones with broad global distribution, can benefit from it just as much. On-demand key distribution, logging, policy-driven automation, and separating key administration from key usage are all examples of best practices. The best practices for centralized key management consist of policy-driven automation, on-demand key distribution, logging, and decoupling key management from key usage.
- ***User Access Control:*** Although user access control is perhaps the most fundamental network security tool, few businesses actually utilize it because of how expensive it can be to operate. This can be disastrous for the big data platform and is already risky at the network level. A policy-based strategy that automates access based on user and role-based settings is necessary for strong user access control. Complex user control levels, such as multiple administrator settings that defend the big data platform against insider attacks, can be managed by policy-driven automation.
- ***Intrusion Detection and Prevention:*** Intrusion detection and prevention systems are security workhorses. This does not make them any less valuable to the big data platform. Big data's value and distributed architecture lend themselves to intrusion attempts. IPS enables security admins to protect the big data platform from intrusion, and should an intrusion succeed, IDS quarantine the intrusion before it does significant damage.
- ***Physical Security:*** Physical security must not be ignored. Build it into the deployment of a big data platform inside a data center, and carefully consider the data center security of the cloud provider. Physical security measures can prevent data center entry to strangers or to staff members who have no job in sensitive areas. Video surveillance and security logs will apply the same thing.

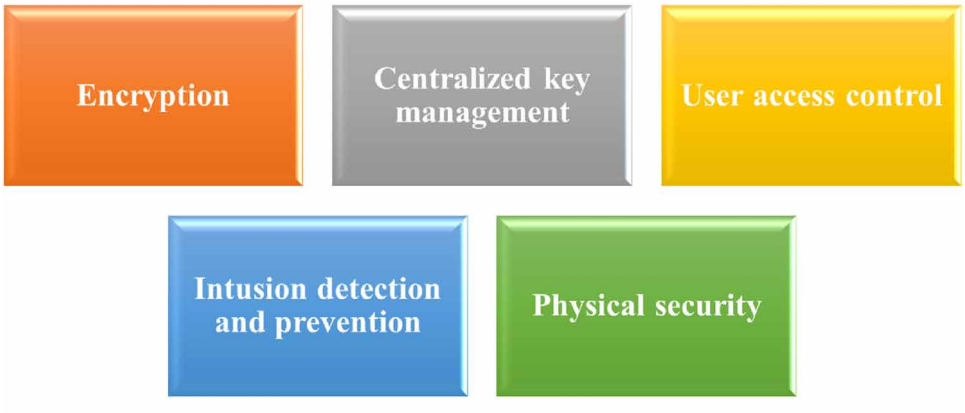
Some businesses using big data securities. A handful of the Big Data security firms' representatives are listed below.

1. ***The Cloudwick CDAP (Cloudwick Data Analytics Platform)*** is a managed security hub that combines security capabilities from several analytics toolkits, conventional IDS and IPS, and machine learning initiatives. CDAP is built on Cloudera's Hadoop software and Intel Xeon processors (Cloudwick, 2022).
2. **IBM**: To keep an eye on Big Data and NoSQL systems, IBM Security Guardium is utilized. It involves finding and identifying sensitive data. Under IBM security, vulnerability assessment, data and file monitoring are also carried out. The questionable access attempts are also hidden, blocked, alerted to, encrypted, and quarantined by Guardium monitoring (IBM, 2020).

3. ***Logtrust:*** Logtrust (recently known as Devo) is partnered with Panda Security in order to provide the ART (Advanced Reporting Tool) and Panda Adaptive Defense. Suspicious digital behavior and internal threats to Big Data systems and networks get automatically reported by the ART. Data from various sources is typically correlated by Panda Adaptive Defense, which is crucial in Big Data environments with numerous nodes and data sources (Devo, 2020).
4. ***Gemalto:*** Gemalto SafeNet safeguards Big Data platforms. Usually, it safeguards the Big Data platforms in the cloud, data center, and virtual environments. The toolbox of security comprises digital signing technologies, data encryption, robust authentication, and cryptographic key security management. Gemalto connects with MongoDB, DataStax, IBM, Cloudera, Zettaset, Hortonworks, and Couchbase, they are leading Big Data providers (Gemalto, 2023).

Figure 3 shows Big Data security technologies.

Figure 3. Big Data security technologies



Implementing Security for Big Data

Whether getting started with big data management and seeking for initial big data security solutions, or being a veteran big data user and require updated protection, here are a few ideas for big data security implementation:

- ***Manage and teach internal users well:*** Inadvertent security blunders by employees present one of the most commonly used security weaknesses to malicious actors. Provide only the bare minimum of data source access to each user depending on their function, implement mobile and company device regulations, and train the staff on security and credential management best practices.
- ***Plan regular security monitoring and audits:*** It's critical to regularly evaluate how the network and data landscape has changed over time, particularly in larger organizations where big data and software are growing almost daily. There are a number of network monitoring technologies and third-party services available on the market, allowing a security personnel immediate access to users and strange behavior. Regular security audits also give a team the chance to evaluate larger-scale concerns before they develop into actual security risks.

- ***Speak with a reliable big data company:*** Companies that provide big data storage, analytics, and managed services typically provide some level of security on their own or in partnership with other companies that do. Talking to provider(s) about the security concerns, regulatory needs, and big data use cases so they may tailor their services to what is needed. The platform used may not have all of the unique features that a business or particular use cases require.

Big Data Security use Cases

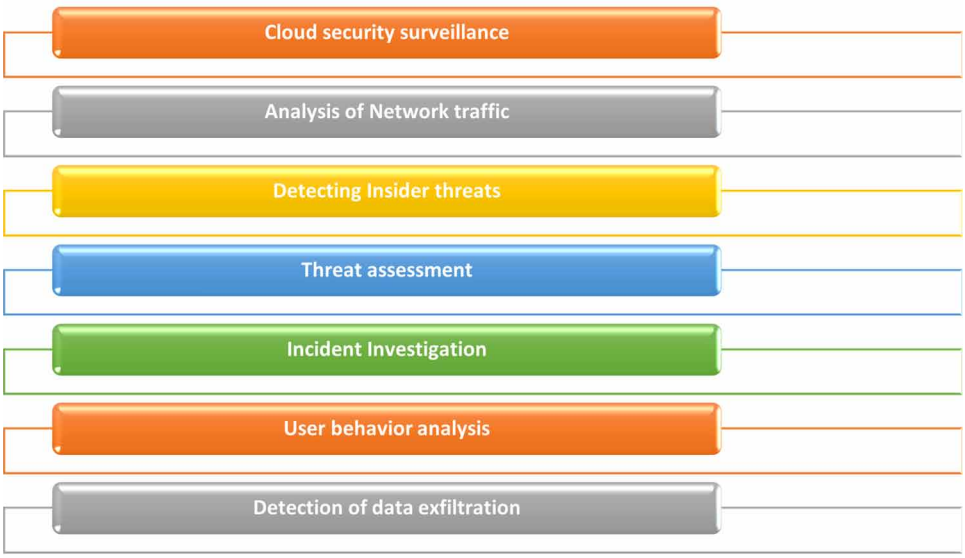
The following topics represent the Big Data security use cases:

1. ***Cloud Security Surveillance.*** In general, cloud computing helps all firms communicate more effectively and profit more. This exchange of information must be protected. Monitoring of cloud applications is provided by big data security tools. This supplies hosts with private data and keeps an eye on cloud infrastructure. Additionally, solutions provide support for a number of pertinent cloud platforms.
2. ***Analysis of Network Traffic.*** The network is always receiving and sending traffic. Maintaining transactional visibility over the network traffic is challenging due to the enormous volume of data being transmitted across the network. A business is able to monitor this network traffic thanks to security analytics. It is employed to create baselines and find abnormalities. Additionally, this aids in monitoring cloud security. It is applied to traffic analysis for cloud infrastructure. It also analyses sensitive data that has been encrypted and illuminates hidden dark areas in infrastructures assuring the channels' proper operation.
3. ***Detecting insider threats.*** A company is just as vulnerable to insider threats as it is to outside ones. Malware attacks cannot do as much harm as an active bad user. A network can be destroyed by an insider threat, however this rarely happens. Security analytics can help firms quickly identify insider risks. Behaviors include odd login times, erratic email usage, and unauthorized database access requests are indicators of this. It occasionally searches for signs that request visibility to outside actors.
4. ***Threat Assessment.*** Generally, the IT Security team mostly engage in threat hunting. They search for potential indicators of dwelling threats and breaches that try to attack the IT infrastructure. Security analytics helps to automate this hunting threat. It acts as an extra set of eyes for your threat hunting efforts. Threats hunting automation can help in detecting malware beaconing activity and thus alerts for its stoppage as soon as possible.
5. ***Incident Investigation.*** Typically, an IT security team would be overwhelmed by the sheer volume of security alerts generated by security information and event management (SIEM) solutions. These continuous alerts can cause more fostering burnout and frustration. Thus to minimize this issue, security analytics automates the incident investigation by providing contextualization to alerts. Thus, a team has more time to prioritize incidents and can deal with potential breach incidents first.
6. ***User Behavior Analysis.*** Organization's users generally interact with an IT infrastructure all the time. Mainly, it is the user's conduct that dictates the success or failure of a cybersecurity. Therefore, there is a necessity for tracking user's behavior. The security analytics monitor the unusual behavior of employees. Thus, it helps to detect an insider threat or a malicious account. It can also detect suspicious patterns by correlating malicious activities. An example of one such renowned security analytics use case is user and entity behavior analytics (UEBA). It helps to provide visibility into the IT environment. Thus, synthesizing user actions from numerous databases into full profiles.

7. ***Detection of Data Exfiltration.*** Data exfiltration is characterized as any unlawful movement of data flowing in and out of any network. Unauthorized data movements can cause theft and leakage of data. Thus, there is a need to protect data from such illegal access. Security analytics enable to detect the data exfiltration through a network. It is primarily used to detect data leakage in encrypted communications.

Figure 4 shows Big Data security use cases.

Figure 4. Big Data security use cases



Big Data Security Implementation in Organizations

One of the most popular security tools is encryption. Encrypted data is hard to decode for hackers. Encrypted data is generally done for the incoming data as well as for outgoing data. Considering the other Big data security tools, then the best one is Firewall. Firewalls are usually used to filter the traffic that enters and leaves servers. Firewall creates strong filters that prevent attacks coming from malicious activities. Business Intelligence (BI) tools and analytics platform is another key to protect vital information of the organization. BI tools are used to build an access system that significantly lowers the likelihood of an attack. Following the recommended procedures for big data security is crucial for the reasons listed below.

- Boosts the security of non-relational data stores.
- Helps to implement endpoint security.
- Ensures the safety of transactions and data storage logs.
- Relies on Big Data Cryptography.
- Utilize specialized services.
- Practice security monitoring and compliance in real-time.
- improves information accessibility and communication.

- facilitates easy resource sharing.
- Increases systems efficiency and robustness.
- Avoids unauthorized access, thus protects and enhances the performance and security of the organization.

Keys to an Effective Big Data Approach

An organization's big data strategy must take into account its business objectives, the data that is currently accessible for usage, and the requirement for new data to support the goals. The following actions are to be taken as the next step:

- putting planned applications and use cases in order;
- determining the required new systems and tools;
- putting together a deployment roadmap; and
- assessing internal talents to determine whether retraining or hiring is necessary.

A data governance program and related data quality management practices must also be prioritized in order to guarantee that sets of large data are clean, consistent, and used appropriately. Aside from prioritizing corporate information needs over available technology, other best practices for handling and interpreting big data include the use of data visualization to speed up data discovery and analysis.

FUTURE RESEARCH DIRECTIONS

There is still a long way to go until Big Data security and privacy issues are resolved, despite some significant advances being taken in that direction. The research community could work hard to create fresh Big Data security and privacy solutions. Data creation (including the Big Data sources - devices), data storage and transportation, data translation and processing, and eventually data utilization all provide research problems. A large capacity and highly distributed architecture exposed to a hostile environment and vulnerable to various types of attacks would be required to sustain this lifespan. More research is required, especially with regard to automatic switching adaption and behavior-based security rules. There are also important research challenges on maintaining end to end data security and privacy. Ensuring that data is never revealed in clear, in particular to non-authorized parties, on any point of the Big Data lifecycle. Moving from data to programs, there are techniques for protecting privacy in browsing, searching, social interactions, and general usage through obfuscation methods.

However, there is more research to be conducted on the processing of encrypted data and privacy protection in the context of both computer programs and web-based systems. More research challenges in the Big Data area include developing techniques to perform a transparent computation over encrypted data with multiple keys, from multiple sources and multiple users. In terms of research it would be challenging to study and develop ways to delegate limited functions over encrypted data, so that third parties can analyze it. All the aspects related with key management, authorization delegation, management of rights, are topics that require further research in this field. When considering safe and private-aware system, trust is everything. For the majority of the cases, especially in the case of Big Data, a reliable environment should be created (healthcare, assisted living, supervisory control and data acquisition (SCADA) systems and many others). It is particularly problematic in terms of research directions how

this environment might be accomplished. A basic security requirement is to trust devices that gather all the data from various sources, as well as applications that can query, process, and extract knowledge from big data. A major study subject for the upcoming years is how trust may be built among end users, devices (IoT), and applications. On what concerns Big Data, these research challenges represent only the tip of the iceberg about the problems that still need to be studied and solved on the development of secure and privacy-aware Big Data ecosystem.

CONCLUSION

There are several challenges in obtaining Big Data analytics and securing Big Data. So when you are hosting the Big Data platform in the cloud, try to take nothing for granted. There is a need to engage closely with the service provider to address these difficulties with strong security SLA (Service Level Agreement). Who is in charge of maintaining the availability and security of this important information? Everyone, or almost everyone, who works for an organization is in charge of protecting the crucial data. Security software, policies, and procedures fall under the purview of IT. Software security helps to protect Big Data deployment against malware and unauthorized access. To protect compliance, which includes protecting the automatic extraction of credit card information, compliance officers must closely collaborate with the IT team. DBAs (Database Administrators) should collaborate closely with the IT team to protect the databases. Securing the Big Data platform against high and low threats will allow firms to perform services well in the long run.

REFERENCES

Agrawal, D., Das, S., & El Abbadi, A. (2011). Big data and cloud computing. *Proceedings of the 14th International Conference on Extending Database Technology - EDBT/ICDT '11*, 530. 10.1145/1951365.1951432

Apache Hadoop. (2015). http://hadoop.apache.org

Apache Mahout. (2015). https://mahout.apache.org/

Apache Storm. (2015). https://storm.apache.org/

Barbacioru, I.C. (2014). *An Illustrative Example Of Application Decision Making Process For Production Consumer Goods*. Academic Press.

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information Communication and Society*, *15*(5), 662–679. doi:10.1080/1369118X.2012.678878

Chen, C. L. P., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, *275*, 314–347. doi:10.1016/j.ins.2014.01.015

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, *19*(2), 171–209. doi:10.100711036-013-0489-0

Cheng, C. K., Chiang, M., & Poor, H. V. (2013). From Technological Networks to Social Networks. *IEEE Journal on Selected Areas in Communications*, *31*(9), 548–572. doi:10.1109/JSAC.2013.SUP.0513049

Cuda. (2015). http://www.nvidia.com/object /cuda_home_new.html

Curtin, R. R., Cline, J. R., Slagle, N. P., March, W. B., Ram, P., Mehta, N. A., & Gray, A. G. (2013). MLPACK: A scalable C++ machine learning library. *Journal of Machine Learning Research*, *14*(80), 1–5.

De Mauro, A., Greco, M., & Grimaldi, M. (2015). What is Big Data? A Consensual Definition and a Review of Key research Topics. *AIP Conference Proceedings*, *1644*(1).

European Parliament. (2021). *Big data: Definition, benefits, challenges* (infographics). https://www.europarl.europa.eu/news/en/headlines/society/20210211STO97614/big-data-definition-benefits-challenges-infographics

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. doi:10.1016/j.ijinfomgt.2014.10.007

Informatica. (2023). https://www.informatica.com/services-and-training/glossary-of-terms/big-data-definition.html

Joseph, R. C., & Johnson, N. A. (2013). Big data and transformational government. *IT Professional*, *15*(6), 43–48. doi:10.1109/MITP.2013.61

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). Big Data: Issues and Challenges Moving Forward. *2013 46th Hawaii International Conference on System Science*, 1530-1605.

Kaur, P., Sharma, M., & Mittal, M. (2018). Big data and machine learning based secure healthcare framework. *Procedia Computer Science*, *132*, 1049–1059. doi:10.1016/j.procs.2018.05.020

Kościelniak, H. & Puto, A. (2015). Big Data in Decision Making Processes of Enterprises. *Procedia Computer Science*, *65*, 1052–1058.

Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with big data. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, *5*(12), 2032–2033. doi:10.14778/2367502.2367572

Marques, J., & Serrão, C. (2013). Improving Content Privacy on Social Networks Using Open Digital Rights Management Solutions. *Procedia Technology*, *9*, 405–410. doi:10.1016/j.protcy.2013.12.045

McKinsey & Company. (2011). *Big data: The next frontier for innovation, competition, and productivity.* McKinsey Global Institute.

Merriam Webster Dictionary. (2023). https://www.merriam-webster.com/dictionary/big%20data#dictionary-entry-1

Millard, S. (2013). *Big Data Brewing Value in Human Capital Management.* Ventana Research. http://stephanmillard.ventanaresearch.com/2013/08/28/big-data-brewing-value-in-humancapital-management

Mosavi, A., & Vaezipour, A. (2013). *Developing Effective Tools for Predictive Analytics and Informed Decisions. Technical Report*. University of Tallinn.

Mygreatlearning. (2020). https://www.mygreatlearning.com/blog/ai-bigdata-case-study/

N-IX. (2020). https://www.n-ix.com/big-data-predictive-analytics-supply-chain-case-study/

Nair, L. R., Shetty, S. D., & Shetty, S. D. (2018). Applying spark based machine learning model on streaming big data for health status prediction. *Computers & Electrical Engineering*, *65*, 393–399. doi:10.1016/j.compeleceng.2017.03.009

NIST Framework. (2019). *NIST big data interoperability framework: Volume 1, definitions*. https://www.nist.gov/publications/nist-big-data-interoperability-framework-volume-1-definitions

Okhravi, H., Hobson, T., Bigelow, D., & Streilein, W. (2014). Finding Focus in the Blur of Moving Target Techniques. *IEEE Security and Privacy*, *12*(2), 16–26. doi:10.1109/MSP.2013.137

Oxford Dictionary. (2023). https://www.oxfordlearnersdictionaries.com/definition/english/big-data

PC Magazine. (2023). https://www.pcmag.com/encyclopedia/term/big-data

Pham, Q. V., Nguyen, D. C., Huynh-The, T., Hwang, W. J., & Pathirana, P. N. (2020). Artificial intelligence (AI) and big data for coronavirus (COVID-19) pandemic: A survey on the state-of-the-arts. *IEEE Access: Practical Innovations, Open Solutions*, *8*, 130820–130839.

Pospiech, M., & Felden, C. (2012). Big data—a state-of-the-art. *Proceedings of the Americas conference on information systems*, 1–23.

Projectpro. (2023). https://www.projectpro.io/article/data-science-case-studies-projects-with-examples-and-solutions/519

Puri, R. (2013). *How Online Retailers Use Predictive Analytics To Improve Your Shopping Experience*. http://blogs.sap.com/innovation/analytics/how-online-retailers-use-predictiveanalytics-to-improve-your-shopping-experience-0108060

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2002). Recommendation systems for large e-commerce: Scalable neighborhood formation using clustering. *Proceedings of the fifth international conference on computer and information technology*, 1.

Shmueli, G., & Koppius, O. (2011). Predictive Analytics in Information Systems Research. *Management Information Systems Quarterly*, *35*(3), 553–572.

Song, Y., Alatorre, G., Mandagere, N., & Singh, A. (2013). *IEEE International Congress on Big Data*.

Systems Plus. (2022). https://www.systems-plus.com/8-case-studies-and-real-world-examples-of-how-big-data-has-helped-keep-on-top-of-competition/

Tan, W., Blake, M., Saleh, I., & Dustdar, S. (2013). Social Network- Sourced Big Data Analytics. *IEEE Internet Computing*, *17*(5), 62–69.

Tawseef, N., Sheriff, I. C., & Qazi, S. (2015). Big data, CEP and IoT: Redefining holistic healthcare information systems and analytics. *International Journal of Engineering Research & Technology (Ahmedabad)*, *4*(1), 1–6.

Techvidvan. (2023). https://techvidvan.com/tutorials/top-10-big-data-case-studies/

Vaishya, R., Javaid, M., Khan, I. H., & Haleem, A. (2020). Artificial intelligence (AI) applications for COVID-19 pandemic. *Diabetes & Metabolic Syndrome*, *14*(4), 337–339.

Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics and big data: A revolution that will transform supply chain design and management. *Journal of Business Logistics*, *34*(2), 77–84.

Zhong, E., Fan, W., Xiao, J. W. L., & Li, Y. (2012). ComSoc: Adaptive Transfer of User Behaviors over Composite Social Network. *18th ACM SIGKDD international conference on Knowledge discovery and data mining*.

Ziora, A.C.L. (2015). The Role of Big Data Solutions in the Management of Organizations. Review of Selected Practical Examples. *Procedia Computer Science*, *65*, 1006–1012.

ADDITIONAL READING

Cloudwick. (2022). https://cloudwick.com/

Devo. (2020). https://www.devo.com/news article/logtrust-is-now-devo/

Gemalto. (2023). https://www.thalesgroup.com/en/markets/digital-identity-and-security

IBM. (2020). https://www.ibm.com/us-en

KEY TERMS AND DEFINITIONS

California Consumer Privacy Act (CCPA): Gives consumers more control over the personal information that businesses collect about them and the CCPA regulations provide guidance on how to implement the law.

Customer Relationship Management (CRM): Customer relationship management is a process in which a business or other organization administers its interactions with customers, typically using data analysis to study large amounts of information.

Cyber Moving Target (MT): Are a collection of technologies that seek to improve security and increase resilience and availability of an application through increasing diversity of software and network paths.

Distributed Denial of Service (DDoS) Attacks: It is a cybercrime in which the attacker floods a server with internet traffic to prevent users from accessing connected online services and sites.

General Data Protection Regulation (GDPR): Is the toughest privacy and security law in the world. It was drafted and passed by the European Union (EU); it imposes obligations onto organizations that collect users' data.

Hadoop Distributed File System (HDFS): Is the primary data storage system used by Hadoop applications. HDFS employs a NameNode and DataNode architecture to implement a distributed file system that provides high-performance access to data across highly scalable Hadoop clusters.

Intrusion Detection System (IDS): An intrusion detection system is a device or software application that monitors a network or systems for malicious activity or policy violations.

Intrusion Protection System (IPS): Is a network security tool (which can be a hardware device or software) that continuously monitors a network for malicious activity and takes action to prevent it, including reporting, blocking, or dropping it, when it does occur.

Non-Relational Databases (NoSQL): Refers to data stores that do not use SQL for queries. Instead, the data stores use other programming languages and constructs to query the data.

Relational Database Management Systems (RDBMS): A relational database is a type of database. It uses a structure that allows us to identify and access data in relation to another piece of data in the database.