WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Secure by Design: Illuminating the Path with Data Science in Cybersecurity

Sankalp Kumar[1], Dr. Febin Prakash[2]

[1] Student of MCA, Department of CS&IT, Jain (Deemed-to-be) University, Bangalore, India
[2] Professor, Department of CS&IT, Jain (Deemed-to-be) University, Bangalore, India
sankalpkumar0229@gmail.com, febin.prakash@jainuniversity.ac.in
DOI: https://doi.org/10.55248/gengpi.5.0324.0705

ABSTRACT

The cybersecurity landscape is changing because of advances in technology, and data science is becoming increasingly important in this shift. This paper gives a brief overview of how data science has evolved and how it's now being used in cloud security, forming what's called cybersecurity data science (CSDS). It emphasizes the importance of using data-driven models to create smart security systems by finding patterns and insights in cybersecurity data. CSDS uses scientific methods, machine learning, and analytics to understand real-world data and improve security measures. It makes traditional cybersecurity methods smarter and more effective. The paper talks about the benefits of CSDS, explaining how to gather relevant data and use analytics to find patterns for better security solutions.

INTRODUCTION

Why is data Important?

Data serves as a cornerstone across diverse sectors due to its multifaceted significance. Firstly, it underpins informed decision-making by offering valuable insights gleaned from analysis, enabling individuals and entities to make well-informed and efficient choices. Furthermore, data aids in identifying and understanding trends and patterns, essential for adapting strategies, forecasting future developments, and maintaining competitiveness.

Moreover, data facilitates performance evaluation by furnishing measurable metrics to track progress towards objectives and gauge the efficacy of endeavors. It also plays a pivotal role in risk management, identifying potential risks through historical and current data analysis and enabling proactive mitigation measures.

The answers to these queries are rooted in the transformative journey that companies have undertaken. While the term "Data Science" is relatively recent, the field has evolved from the role of statisticians. In the pre-Data Science era, statisticians, specializing in qualitative data analysis, were employed by companies to assess overall performance and sales. The convergence of computer science with statistics, facilitated by computing processes, cloud storage, and analytical tools, gave rise to the field of Data Science.

Early data analytics primarily focused on surveys and addressing public issues. For instance, a survey about the number of children in a district could inform decisions on school development. The integration of computers simplified decision-making, enabling the resolution of more intricate statistical problems. As data proliferated, companies recognized its value, leading to the development of products aimed at enhancing customer experiences. Industries sought experts capable of harnessing the potential embedded in data, utilizing it to make informed business decisions, maximize profits, and understand customer behavior based on purchasing patterns. Data became instrumental in refining revenue models and improving product quality.

In the context of products, data is akin to electricity for household gadgets. It is indispensable for engineering products tailored to users' needs, driving the product and rendering it usable. Comparably, a data scientist functions as a sculptor, meticulously shaping data to derive meaningful outcomes. Although the task may be intricate, a proficient data scientist with the right expertise can deliver impactful results.

Importance of data science

Data possesses transformative powers. Industries rely on data for making informed decisions, and Data Science acts as the alchemy that transforms raw data into valuable insights. Hence, there is a crucial need for Data Science in various sectors. A Data Scientist is akin to a sorcerer, wielding the ability to extract meaningful information from any dataset. A proficient Data Scientist guides a company by steering through data to uncover pertinent insights. The expertise of a skilled Data Scientist is essential for steering the company in the right direction with robust, data-driven decisions. Proficient in the realms of Statistics and Computer Science, the Data Scientist employs analytical prowess to address and solve intricate business challenges.

Purpose of Data - Centric Industries

Industries that are centered around data prioritize its strategic utilization as a fundamental component in their operational and decision-making procedures.

Data Science for Enriching Lives

Data Science enhances lives by utilizing data to create positive and meaningful impacts across diverse fields. In the healthcare sector, it assumes a crucial role in personalized medicine, tailoring garments based on individual genetic information and health data, resulting in more effective and less invasive interventions. Within education, data-driven insights facilitate personalized learning experiences, identifying real-time strengths and weaknesses and promoting adaptive learning approaches.

In the business and finance realm, Data Science streamlines decision-making processes, allowing companies to fine-tune strategies, elevate customer experiences, and identify and prevent fraud. Smart cities harness data for improved urban living, enhancing transportation systems, energy efficiency, and public services. In social sciences, Data Science analyzes demographic and societal data, assisting policymakers in crafting targeted interventions and addressing social challenges. Ultimately, Data Science enhances lives by extracting valuable insights from data, fostering innovation, and contributing to progress in healthcare, education, business, and societal well-being. It empowers individuals and organizations to make informed decisions, instigating positive change and elevating overall quality of life.

Data Science for Innovation

Data Science is a catalyst for innovation, revolutionizing how industries operate and driving transformative changes. By harnessing the power of data, organizations can uncover hidden patterns, trends, and insights that spark innovative ideas. In product development, Data Science informs the creation of cutting-edge solutions by analyzing customer feedback, market trends, and user behavior.

In healthcare, Data Science fuels innovation through personalized medicine, where treatments are tailored to individual genetic profiles and health data. The ability to analyze massive datasets also enables researchers to identify potential breakthroughs, accelerating drug discovery processes and advancing medical knowledge.

In business, Data Science optimizes operations, enhances decision-making, and identifies opportunities for growth and efficiency. From predictive analytics that anticipate market trends to recommendation systems that personalize user experiences, Data Science fosters innovation by leveraging data-driven intelligence. Smart cities leverage Data Science to innovate urban living, improving infrastructure, transportation, and resource management. In research, Data Science accelerates scientific discoveries by analyzing complex datasets, uncovering correlations, and guiding experiments.

Data Science for Better Marketing

Leveraging data science in marketing is pivotal for success in today's data-centric environment. With advanced analytics and machine learning algorithms, data science empowers marketers to gain profound insights into consumer behavior, preferences, and trends. This wealth of information allows for the creation of targeted and personalized marketing campaigns, ensuring that messages resonate with specific audiences. Moreover, data science facilitates the development of accurate prediction models, enabling marketers to anticipate market trends, optimize pricing strategies, and identify potential customer segments. Through data-driven decision-making, businesses can refine their marketing approaches, allocate resources more efficiently, and ultimately enhance their return on investment.

Data science acts as a driving force for innovation in marketing, providing the tools necessary to connect with audiences on a deeper level, foster customer engagement, and achieve business growth in an increasingly competitive marketplace.

Role of Data Science in Cyber Security

At its core, the fundamental essence of data science revolves around comprehension. This involves the examination, manipulation, and extraction of valuable insights from a given set of information. While the concept and practice have been in existence for several decades, initially as a subset of computer science, it has evolved into an independent field. As a result, individuals with an interest in the subject can now pursue dedicated studies and majors in data science.

One contemporary application of data science is evident in the realm of cybersecurity. Although it may initially seem unconventional to explore data science to enhance cybersecurity, it is a logical and practical choice. The synergy lies in the fact that data science methodologies are instrumental in fortifying cybersecurity measures, providing a robust and effective approach to address and mitigate potential threats.

Relation between Big Data and Cyber Security

The correlation between Big Data and cybersecurity is essential in tackling the escalating challenges of securing digital assets and networks. Big Data technologies play a pivotal role in managing the sheer volume and intricacy of data generated within the cybersecurity domain. As organizations grapple with a constant stream of security-related information, such as logs, network traffic, and system events, the capabilities of Big Data processing become indispensable. These technologies enable real-time monitoring and analysis, enabling cybersecurity professionals to promptly detect anomalies, patterns, and potential threats. Additionally, the scalability and storage features of Big Data solutions are crucial for retaining extensive historical data, facilitating forensic analysis, and ensuring compliance. The integration of machine learning and artificial intelligence with Big Data contributes to adaptive security measures by learning from past incidents and adapting to emerging cyber threats. In summary, the interplay between Big Data and cybersecurity empowers

organizations to not only handle the voluminous data but also derive actionable insights, bolster incident response capabilities, and strengthen defenses against the ever-evolving landscape of cyber threats.

Predictive and Active Intrusion Detection System:

A Predictive Intrusion Detection System and an Active Intrusion Detection System offer distinct but complementary approaches to strengthening cybersecurity. The Predictive Intrusion Detection System employs machine learning algorithms and statistical models to analyze historical data and detect patterns indicative of future security threats. By establishing a baseline of normal network behavior, it can predict and notify security personnel about potential intrusion attempts before they occur, providing a proactive defense against evolving cyber threats.

In contrast, the Active Intrusion Detection System takes a proactive stance in cybersecurity. Instead of solely identifying threats, it actively responds in real-time with automated actions to mitigate the impact of intrusions. These responses include blocking malicious traffic, isolating compromised systems, or triggering alerts for further investigation. By promptly addressing identified threats, the Active Intrusion Detection System aims to prevent or minimize the damage caused by malicious activities, thereby enhancing the overall resilience of the network.

Together, these systems contribute to a comprehensive cybersecurity strategy by integrating predictive analytics with active responses. While the Predictive Intrusion Detection System anticipates and warns of potential threats, the Active Intrusion Detection System takes swift actions to neutralize and mitigate the impact of those threats, creating a layered defense against the dynamic landscape of cyber threats.

Protecting valuable information:

Safeguarding valuable information is a paramount concern in the face of potential data attacks. The risk of losing highly valuable data and information poses a significant threat to an organization's integrity. Implementing robust security measures, such as intricate signatures or encryption, serves as a deterrent against unauthorized access to datasets. Integrating Data Science into security protocols becomes instrumental in creating formidable defenses. For instance, through the analysis of historical cyber-attack data, one can devise algorithms capable of identifying the most frequently targeted segments of data. In essence, the application of Data Science, facilitated by efficient Analytics Systems, not only fortifies the cybersecurity industry but also empowers IT professionals to develop more effective, defensive, and proactive measures to thwart cyber-attacks.

Data Science answers to all cyber-security challenges

The continuous evolution of data science and machine learning is increasingly relevant in the realm of data security, with AI in cybersecurity projected to reach nearly $35 billion by 2025. Data scientists contribute their expertise to fortify cybersecurity efforts by thwarting attacks and detecting suspicious activities. Serving as technical experts, problem gatherers, analysts, and adept interpreters, data scientists facilitate problem-solving in diverse capacities. By applying data science knowledge, coders and programmers enhance their capabilities to develop more effective programs against cyber threats.

In the cybersecurity industry, there is a constant demand for technical resources, emphasizing the need for individuals who are not only proficient in coding but also possess sharp problem-solving skills. This makes them highly sought after, often commanding lucrative job offers. Moreover, the staggering financial losses incurred through data breaches each year underscore the critical importance of robust cybersecurity measures.

Data science proves instrumental in cybersecurity programs by focusing on the identification of threats, prevention of intrusions and attacks, accurate detection of malware and spam, and the mitigation of fraud risks. Machine learning and data science enable improved threat identification by utilizing extensive datasets for deep learning and training. In the case of malware and spam detection, this approach aims to reduce false positives, optimizing time and resources.

Addressing intrusions and attacks, data science plays a pivotal role in recognizing anomalies in user behavior that may signal an intruder's presence. By identifying these abnormalities early on, preventive measures can be implemented to curtail the severity of potential intrusions. This approach is particularly relevant in scenarios like Ransomware attacks, which have seen a 37 percent increase in cases last year.

In the context of fraud prevention, data science follows a similar methodology. Analyzing samples from datasets, anomalies in activities such as credit card purchases are detected, providing valuable insights to identify and combat fraudulent behavior. Overall, data science serves as a powerful tool to connect the dots between seemingly minor abnormalities, allowing for a comprehensive understanding of potential cybersecurity threats.

Challenges to face

Not relying on sequences:

A significant advantage of employing data science in cybersecurity lies in its ability to utilize extensive datasets, moving away from reliance on "lab-based" sequences. Unlike traditional cybersecurity programs built on pre-defined sequences of events, data science allows for the analysis of larger samples of real-world data. This departure from predetermined sequences is crucial because hackers seldom adhere to established "rules" or patterns.

When developing a program to identify threats, it becomes essential to evaluate authentic data derived from real users. This approach enables the recognition of genuine normal behavior, a critical prerequisite for effectively identifying abnormal behavior. In essence, the use of data science ensures a more realistic and adaptable approach to cybersecurity, taking into account the dynamic and unpredictable nature of cyber threats.

Focusing on the abnormalities

Not every slightly unusual behavior holds relevance in the realm of cybersecurity. To effectively address this, understanding the reasons behind such behaviors becomes imperative to minimize false positives.

In the cybersecurity landscape, deviations from what is considered normal activity are inevitable. For instance, individuals may travel to different countries and log in from various locations, use different devices for logging in, or unexpectedly make purchasing decisions that deviate from their established history. The key lies in recognizing the context surrounding such behaviors, as the same type of activity can have different implications based on the broader circumstances. It is crucial to acknowledge that there may be extraneous noise in the data that is not pertinent, potentially leading to the generation of numerous false positives.

The importance of applying data science in the realm of cybersecurity.

In the ever-evolving landscape of cybersecurity, the application of data science emerges as a critical and indispensable tool, playing a pivotal role in fortifying defenses, identifying threats, and staying one step ahead of cyber adversaries. The significance of employing data science in cybersecurity extends across various dimensions, encompassing proactive threat detection, adaptive risk mitigation, and the continuous enhancement of security measures.

One of the primary advantages of leveraging data science in cybersecurity lies in its ability to handle vast volumes of data. Traditional methods often struggle to analyze the sheer magnitude of information generated in the digital realm. Data science, equipped with advanced algorithms and machine learning techniques, enables the processing of large datasets at unprecedented speeds. This not only facilitates real-time threat detection but also allows for a more comprehensive understanding of patterns and anomalies within the data.

Moreover, data science provides a departure from conventional, rule-based approaches to cybersecurity. Traditional security systems often rely on predefined rules and signatures, rendering them susceptible to evasive tactics employed by modern cyber threats. In contrast, data science enables the development of adaptive models that learn and evolve over time. This dynamic approach is crucial in addressing the ever-changing tactics of cyber adversaries who constantly seek new ways to infiltrate systems.

A fundamental aspect of cybersecurity is the ability to distinguish between normal and abnormal behavior. Here, data science excels by focusing on anomalies that may signify potential threats. However, it is essential to note that not every deviation from the norm is a cause for concern. Context is key, and data science allows for a nuanced understanding of behaviors, reducing the risk of false positives. For instance, a sudden change in login location or device may be entirely benign, such as when users travel or upgrade their devices. By discerning the context surrounding such deviations, data science contributes to more accurate threat identification. The predictive capabilities of data science also elevate cybersecurity defenses to a proactive stance. Through the analysis of historical data and patterns, machine learning algorithms can anticipate potential threats before they materialize. This proactive approach empowers organizations to implement preemptive measures, significantly reducing the risk of security breaches. Additionally, the continuous learning nature of data science models ensures adaptability to emerging threats, reinforcing the resilience of cybersecurity systems.

The significance of data science in cybersecurity is further underscored by the increasing complexity of cyber threats. As adversaries employ sophisticated techniques and exploit vulnerabilities across diverse attack vectors, traditional methods alone prove insufficient. Data science, with its ability to correlate and analyze multifaceted datasets, provides a holistic view of potential threats. This comprehensive analysis enables security professionals to identify subtle indicators and uncover hidden patterns that may elude conventional security measures.

In an era where the cybersecurity landscape is marked by rapid advancements and persistent challenges, the integration of data science stands as a linchpin for effective defense strategies. The fusion of advanced analytics, machine learning, and artificial intelligence not only enhances the efficiency of threat detection but also empowers organizations to stay agile in the face of evolving cyber risks. The importance of applying data science in the realm of cybersecurity is not merely a technological evolution but a strategic imperative to safeguard digital assets and uphold the integrity of the interconnected world.

The concern regarding AI-driven cyber attacks

Traditional malware is typically identifiable and preventable through the use of signatures, which leave traces that are collected and distributed as indicators of compromise (iOS) to antivirus engines. These signatures are then utilized to scan files entering a network or computer, promptly quarantining or deleting any matches. However, traditional antivirus methods encounter challenges such as delays, scalability issues, and the continuous growth of signature lists, impacting storage efficiency.

Unlike traditional approaches, antivirus systems employing artificial intelligence (AI) focus on detecting abnormal behavior exhibited by programs instead of relying on signatures. This shift allows for the identification of zero-day exploits and previously unknown malware by recognizing actions that deviate from the standard operation of a computer. This proactive approach addresses the delay associated with signature availability and improves overall detection efficiency. However, the landscape changes when considering AI-driven cyber-attacks. While AI-based antivirus systems excel in distinguishing normal glitches from common attacks, the scenario becomes more precarious with AI-driven attacks. The challenge lies in the scarcity of datasets available to train systems to differentiate between routine malware attacks and those orchestrated by AI. This scarcity impedes the ability to effectively prepare and defend against AI-driven cyber threats.

In the realm of Cybersecurity Data Science, the focus shifts towards addressing this evolving challenge. The intricacies of AI-driven attacks necessitate innovative data science approaches to enhance the understanding of malicious behavior. Cybersecurity data scientists work on developing models that can adapt and learn from emerging threats, leveraging the limited datasets available to anticipate and counteract the sophistication of AI-driven cyber-attacks. The fusion of data science and cybersecurity becomes crucial in navigating this dynamic landscape, ensuring that defense mechanisms evolve in tandem with the ever-changing nature of cyber threats.

Cybersecurity Data Science

The global landscape is experiencing a significant transformation driven by the widespread adoption of data science, leading many industries to embrace a data-centric approach. This shift has profound implications for the future of intelligent cybersecurity systems and services, as the essence of security lies in effectively managing data. Security professionals traditionally relied on methods like file hashes, custom rules, or manual heuristics to detect cyber threats, without leveraging data science techniques.

However, in recent times, there has been a notable shift towards integrating data science into cybersecurity practices. This recognition stems from the ability of data science to convert raw data into actionable insights. This shift involves various data-driven tasks, including data engineering for efficient data gathering and analysis. An essential aspect of this process is reducing data volume by filtering significant and relevant information for further analysis. Additionally, there is a focus on discovery and detection, aiming to extract insights and incident patterns from the accumulated data. Automated models are being developed to create data-driven intelligent security systems, while targeted security alerts are generated based on discovered knowledge, to minimize false alerts. In the realm of data-driven decision-making, behavioral analysis assumes a significant role within the domain of cybersecurity. As industries increasingly adopt data science methodologies, the emphasis on behavioral analysis becomes instrumental in enhancing the effectiveness of cyber threat detection and response mechanisms. Therefore, the concept of cybersecurity data science involves integrating the methodologies and techniques of both data science and machine learning, alongside the behavioral analytics of diverse security incidents. The combination of these technologies has led to the emergence of the term "cybersecurity data science," signifying the process of collecting a significant amount of security event data from various sources and subjecting it to analysis using machine learning technologies. This analytical approach aims to identify security risks or attacks by revealing valuable insights or recognizing the most recent data-driven patterns.

It is crucial to acknowledge that cybersecurity data science extends beyond the mere implementation of a set of machine learning algorithms. Instead, it embodies a systematic approach crafted to support security professionals or analysts in scaling and automating their security activities with intelligence and efficiency. Consequently, a formal definition of cybersecurity data science can be formulated as follows: "Cybersecurity data science constitutes a research or operational domain positioned at the intersection of cybersecurity, data science, and machine learning or artificial intelligence. Focused primarily on security data, it utilizes machine learning methods, seeks to quantify cyber-risks or incidents, and advocates for inferential techniques to analyze behavioral patterns within security data. Furthermore, it underscores the generation of security response alerts and ultimately aims to optimize cybersecurity solutions, leading to the development of automated and intelligent cybersecurity systems."

Anticipated challenges in the field of Cybersecurity Data Science include

Data Privacy Concerns:

As the volume of data collected for cybersecurity purposes increases, there are growing concerns about preserving the privacy of individuals. Striking a balance between effective threat detection and respecting privacy regulations is a critical challenge.

Lack of Quality Datasets:

Building robust machine learning models relies heavily on high-quality, diverse datasets. However, in the realm of cybersecurity, obtaining labeled datasets that accurately represent evolving threats can be challenging. The scarcity of such data hinders the training of effective models.

Sophistication of Cyber Threats:

Cyber adversaries continually evolve their tactics, techniques, and procedures. Keeping pace with the sophistication of modern cyber threats poses a significant challenge for cybersecurity data scientists. Adaptive and dynamic models are required to effectively counter these evolving threats.

Interdisciplinary Skill Gap:

Cybersecurity Data Science requires a blend of expertise in cybersecurity, data science, and machine learning. Bridging the interdisciplinary skill gap and fostering collaboration among professionals with diverse backgrounds remains a challenge.

Explain ability and Interpretability:

The interpretability of machine learning models in cybersecurity is crucial for gaining trust and understanding the decisions made by these models. Developing models that are not only accurate but also explainable is a persistent challenge.

Adversarial Attacks:

Adversarial attacks involve manipulating input data to deceive machine learning models. In cybersecurity, adversaries may intentionally design attacks to bypass detection models, highlighting the need for robust models that can withstand such manipulation.

Real-time Analysis and Response:

Cyber threats often unfold in real-time, requiring swift analysis and response. Ensuring that cybersecurity data science models can operate in real-time to detect and mitigate threats promptly is an ongoing challenge.

Resource Constraints:

Deploying and maintaining sophisticated cybersecurity data science solutions can be resource-intensive. Organizations face challenges in terms of infrastructure, computational power, and skilled personnel, particularly for smaller entities with limited resources.

Global Regulatory Compliance:

The field of cybersecurity operates within a complex web of global regulations and compliance standards. Ensuring that cybersecurity data science practices align with these regulations, such as GDPR or HIPAA, presents a persistent challenge.

Continuous Learning and Adaptation:

The threat landscape is dynamic, with new attack vectors emerging regularly. Cybersecurity data science models need to be continuously updated and adapted to address emerging threats, requiring a robust framework for continuous learning. Navigating these challenges demands ongoing research, collaboration across disciplines, and a commitment to staying ahead of the curve in both technological advancements and adversarial tactics.

Conclusion

This paper delves into the combination of cybersecurity, data science, and machine learning, exploring how cybersecurity data science enhances intelligent decision-making in data-driven security systems and services. It examines the impact of cybersecurity data science on security data, focusing on insights derived from security incidents and their underlying datasets. The discussion reviews the current state of security incident data and related services, as well as the role of machine learning techniques in cybersecurity.

The paper highlights a gap in research, noting a predominant focus on traditional security solutions over machine learning-based systems. It thoroughly evaluates common techniques in terms of their relevance to security research and aims to provide an overview of conceptualization, understanding, modeling, and considerations in cybersecurity data science. Key issues in security analysis are identified and discussed, laying the groundwork for future research directions.

Drawing from existing knowledge, the paper introduces a comprehensive multi-layered framework for cybersecurity data science based on machine learning techniques. This framework covers phases such as security data collection, preparation, machine learning-based security modeling, and adaptation to incremental learning and dynamism. Ultimately, it aims to contribute to the development of intelligent cybersecurity systems and services. The paper underscores the significance of gleaning insights from security data, presenting a research framework with a keen focus on developing data-driven intelligent security solutions. Its in-depth analysis and discourse carry implications for both security researchers and practitioners

REFERENCES

[1] W. v.d. Aalst, "Data scientist: The engineer of the future", *Proc. I-ESA Conf.*, vol. 7, pp. 13-28, 2014.

[2] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, et al., *Big data: The Next Frontier for Innovation Competition and Productivity*, 2011.

[3] W. v.d. Aalst, Process Mining: Discovery Conformance and Enhancement of Business Processes, Berlin, Germany:Springer-Verlag, 2011.

[4] W. v.d. Aalst, *Proc. IEEE Int. Enterprise Distrib. Object Comput. Conf.*, pp. 1-1, 2014.

[5] Big Data: The Next Frontier for Innovation Competition and Productivity, 2011.

[6] B. F. Jones, S. Wuchty and B. Uzzi, "Multi-University Research Teams: Shifting Impact Geography and Stratification in Science", *Science*, vol. 322, pp. 1259-1262, 2008.

[7] C. L. Philip, Q. Chen, and C. Y. Zhang, "Data-intensive applications challenges techniques and technologies: A survey on big data", *Information Sciences*, vol. 275, pp. 314-347, 2014.

[8] https://en.wikipedia.org/wiki/Data_science.

[9] J. Bollen, H. Van, de Sompel, A. Hagberg, R. Chute, M. A. Rodriguez, et al., "Clickstream Data Yields High-Resolution Maps of Science", *PLoS ONE 4*, pp. 1-11, 2009.

[10] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters", *Commun.ACM*, vol. 51, no. 1, pp. 107-113, Jan 2008.

[11] J. Manyika, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A. Byers, from Big data: The Next Frontier for Innovation Competition and Productivity, 2011.

[12] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, A. Konwinski, G. Lee, et al., "A view of cloud computing", *Commun. ACM*, vol. 53, no. 4, pp. 50-58, Apr 2010.

[13] M. Hilbert and P. Lopez, "The world's technological capacity to store communicate and compute information", *Science*, vol. 332, no. 6025, pp. 60-65, 2011.

[14] M. K. Kakhani, S. Kakhani and S. R. Biradar, Research issues in big data analytics International Journal of Application or Innovation in Engineering Management, vol. 2, no. 8, pp. 228-232, 2015.

[15] M. M. Waldrop, "Complexity: The Emerging Science at the Edge of Order and Chaos", *Simon Schuster*, 1992.

[16] P. Chapman, J. Clinton, R. Kerber, C. Shearer and R. Wirth, "CRISP-DM 1.0: Step-by-step data mining guide", *The CRISP-DM Consortium*, 2000.

[17] S. Wuchty, B. F. Jones and B. Uzzi, "The Increasing Dominance of Teams in Production of Knowledge", *Science*, vol. 316, pp. 1038-1039, 2007.