A REVIEW ON BIG DATA INTEGRATION's DIFFICULTIES WITH AI

Dr. G. Thippanna[1]
[1]Professor, Dept. of MCA, Ashoka Women's Engineering College.
gt.pana2012@gmail.com

Dr. D. William Albert[2]
[2]Professor, Dept. of CSE, Ashoka Women's Engineering College.
dr.albertdwgtl@gmail.com

E. Ramachandra[2]
[2]Asst. Professor, Dept. of MCA, Ashoka Women's Engineering College.
ramchandra535@gmail.com

Abstract

In this review article explored massive amounts of data from big data are expected to revolutionize artificial intelligence (AI). The idea of Next-Gen Big Data Intelligence is the key component that powers AI platforms and helps to release its enormous and untapped potential for mass production and targeted consumption, which will have a significant impact on our society. When offering Big Data Intelligence, such game-changing technology must do so with openness, justice, trust, and reduced prejudice. Sadly, the current Big Data technologies veer off this course, where AI platforms typically operate within centralized proprietary organizations that manage, and control them, exposing critical issues of AI algorithmic and data bases, along with insidious and pervasive reinforcement of discriminatory societal practices. When utilizing unprocessed, noisy, or unintentionally biased data to train AI models, this is particularly troublesome.

***Keywords*: -** AI, Big Data, Data Processing, Big Data Intelligence, Models, Process, Applications.

1. Introduction to AI

AI with big data refers to the use of artificial intelligence (AI) [1] techniques in analyzing and extracting insights from large and complex datasets. The combination of AI and big data allows organizations to uncover patterns, make predictions, and gain valuable insights that can inform decision-making and improve various processes.

Here's a high-level overview of how AI with big data typically works:

- **Data collection:** The first step involves gathering and storing large volumes of structured and unstructured data from various sources, such as databases, sensors, social media, websites, and more. This data can include text, images, videos, numerical values, and other types of information.
- **Data preprocessing:** Once the data is collected, it needs to be cleaned, organized, and prepared for analysis. This process involves removing noise and irrelevant information, handling missing values, standardizing formats, and transforming the data into a suitable format for AI algorithms.
- **AI algorithm selection:** Depending on the nature of the problem and the type of insights sought, different AI algorithms can be applied. These algorithms can include machine learning techniques like classification, regression, clustering, and deep learning methods such as neural networks.
- **Training the AI model:** In this stage, the AI model is trained using the prepared dataset. For supervised learning, the model is trained on labeled data, where it learns patterns and relationships between input variables and target variables. Unsupervised learning can also be used to discover hidden patterns or groupings in the data without predefined labels.
- **Model evaluation and refinement:** After training, the AI model is evaluated to assess its performance and accuracy. Evaluation metrics depend on the specific problem, such as accuracy, precision, recall, or F1 score. If the model's performance is not satisfactory, it can be refined by adjusting hyper parameters, changing the architecture, or collecting more data.
- **Data analysis and insights:** Once the AI model is trained and validated, it can be used to analyze the big data at scale. The model applies its learned knowledge to the new data, identifying patterns, making

predictions, detecting anomalies, or clustering similar data points. These insights can be used for decision-making, process optimization, personalized recommendations, fraud detection, and many other applications.

- **Continuous learning and improvement:** AI with big data is an iterative process. As new data becomes available, the AI model can be updated and retrained to adapt to changing patterns or improve its performance. This continuous learning allows the model to evolve and provide more accurate insights over time.

It's important to note that AI with big data also involves considerations of data privacy, security, ethical use of data, and compliance with regulations to ensure responsible and beneficial outcomes.

1.1. Process of AI

The process of AI (Artificial Intelligence) involves several key steps. Here's a general overview of the AI process [3]:

- **Problem Definition:** The first step is to clearly define the problem or task that the AI system aims to solve or accomplish. This involves understanding the problem domain, identifying the goals, and determining the specific requirements.
- **Data Collection:** AI systems require large amounts of relevant data to learn and make accurate predictions or decisions. This step involves gathering and collecting data from various sources, such as databases, sensors, APIs, or the internet. The quality and diversity of the data are crucial for the performance of the AI system.
- **Data Preprocessing:** Raw data is often unstructured or contains noise, missing values, or inconsistencies. Data preprocessing involves cleaning, transforming, and normalizing the data to make it suitable for AI algorithms. This step may include tasks like data cleaning, feature selection, normalization, and handling missing data.
- **Model Selection and Training:** In this step, an appropriate AI model or algorithm is chosen based on the problem type and the available data. Examples of AI models [4] include neural networks, decision trees, support vector machines, and Bayesian networks. The selected model is then trained on the preprocessed data by adjusting its internal parameters to minimize errors or optimize a specific objective function.
- **Model Evaluation:** After training the AI model, it needs to be evaluated to assess its performance and generalization capabilities. This involves using a separate set of data called a validation or test set to measure metrics like accuracy, precision, recall, or F1 score. The model may go through multiple iterations of training and evaluation until satisfactory performance is achieved.
- **Model Deployment:** Once the AI model has been trained and evaluated, it can be deployed for real-world applications. This involves integrating the model into the target environment or system where it will be used. The deployment process may include considerations such as scalability, efficiency, security, and user interface design.
- **Monitoring and Maintenance:** AI systems require ongoing monitoring to ensure they continue to perform well over time. This includes tracking the system's performance, detecting and resolving any issues or errors, and updating the model or algorithm as needed. Maintenance also involves keeping the data up to date and relevant to maintain the system's accuracy and effectiveness.
- **Iteration and Improvement:** AI is an iterative process, and there is always room for improvement. Feedback from users, new data, or changing requirements can lead to updates and enhancements to the AI system [5]. This may involve retraining the model with additional data or refining the algorithms to achieve better results.

It's important to note that the specific details and techniques involved in each step can vary depending on the particular AI problem, the available resources, and the chosen AI approach.

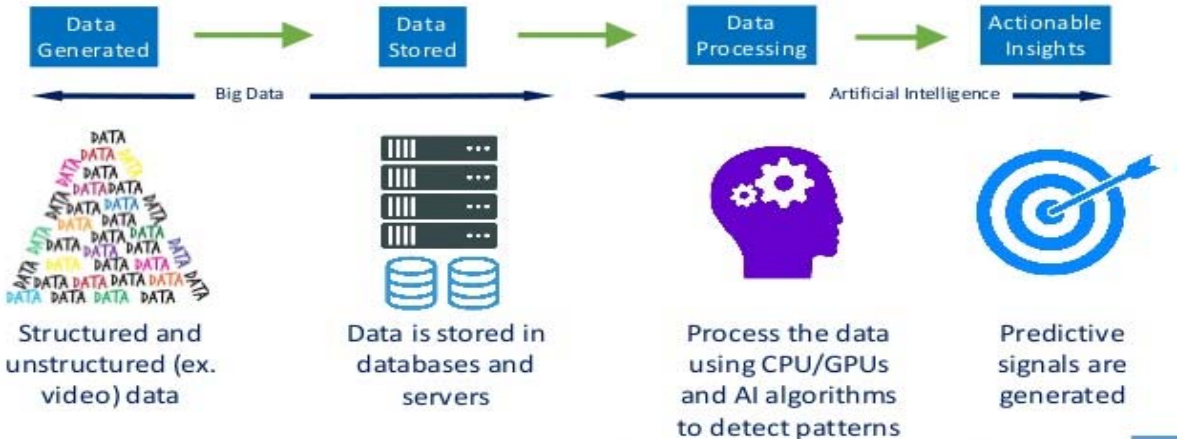

Diagram 1: Diagrams roughly explaining the process of AI

1.2. Architecture of AI

The architecture of an artificial intelligence (AI) [6-9] system refers to its underlying design and structure that enables it to perform specific tasks and functions. AI architectures can vary depending on the specific application and approach, but there are some common components and concepts found in many AI systems. Here are some key elements typically involved in the architecture of an AI system:

a. **Data:** AI systems rely on large amounts of data to learn and make predictions or decisions. This data can be labeled or unlabeled and is used for training, validation, and testing purposes. The quality and quantity of data are crucial factors that can significantly impact the performance of an AI system.
b. **Preprocessing:** Before the data is used for training or inference, it often undergoes preprocessing steps. This can include data cleaning, normalization, feature extraction, and transformation to ensure it is in a suitable format for the AI algorithms [10].
c. **Model:** The model is a core component of an AI system. It represents the learned knowledge and the ability to make predictions or perform specific tasks. The model can be based on various machine learning techniques such as supervised learning, unsupervised learning, reinforcement learning, or a combination of these. Deep learning models, such as neural networks, have gained significant popularity in recent years due to their ability to learn complex patterns and representations.
d. **Training:** In the training phase, the AI system learns from the provided data to optimize its model parameters. This involves feeding the training data into the model, comparing its predictions to the known ground truth, and adjusting the model's parameters through an optimization process (e.g., gradient descent) to minimize the prediction errors.
e. **Evaluation and Validation:** Once the model is trained, it needs to be evaluated to assess its performance. This is typically done using separate validation or test datasets that were not used during the training phase. Evaluation metrics such as accuracy, precision, recall, or F1 score can be used to measure the model's performance and identify areas for improvement.
f. **Inference:** After the model is trained and validated, it can be deployed for inference or prediction on new, unseen data. Inference involves using the trained model to make predictions or decisions based on the input data. The AI system takes the input, processes it through the model, and produces the desired output or action.
g. **Deployment:** The deployment phase involves integrating the AI system into a production environment where it can be used to perform real-world tasks. This can involve setting up the necessary infrastructure, optimizing the model for efficient inference, and ensuring the system's reliability, scalability, and security.
h. **Feedback Loop:** AI systems often have a feedback loop to continuously improve their performance. This can involve monitoring the system's output, collecting feedback or new data from users, retraining the model periodically with updated data, and iterating on the architecture to address any limitations or shortcomings.

It's important to note that AI architectures can vary significantly depending on the specific application and domain. Different techniques and approaches may be employed based on the problem being solved, available data, computational resources, and other factors.

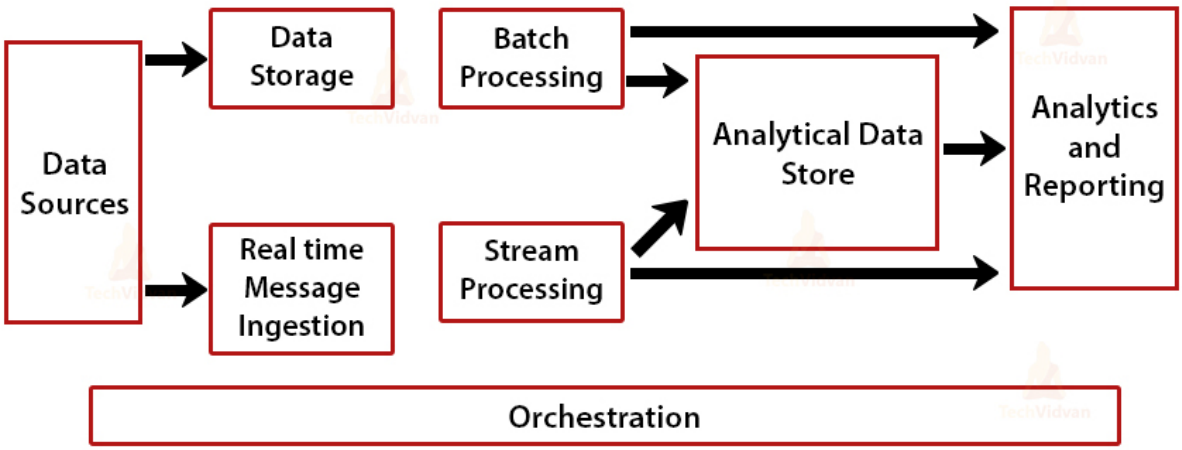

Diagram 2: Big Data Architecture

1.3. Applications of AI

Artificial Intelligence (AI) has numerous applications across various domains. Here are some prominent applications of AI:

- **Virtual Assistants:** AI-powered virtual assistants like Siri, Google Assistant, and Alexa provide voice-based interaction and perform tasks such as answering questions, setting reminders, and controlling smart devices.
- **Recommendation Systems:** AI algorithms analyze user preferences and behavior to provide personalized recommendations in various areas like e-commerce, streaming services, social media, and content platforms.
- **Natural Language Processing (NLP):** AI techniques enable computers to understand and generate human language. NLP is used in applications like chat bots, language translation, sentiment analysis, and voice recognition.
- **Autonomous Vehicles:** AI plays a vital role in self-driving cars. It involves computer vision, sensor fusion, and machine learning to perceive the environment, make decisions, and navigate without human intervention.
- **Healthcare:** AI assists in medical diagnosis, drug discovery, and personalized treatment plans. It can analyze vast amounts of patient data, assist in radiology and pathology, and identify patterns that may be difficult for human doctors to detect.
- **Financial Services:** AI algorithms are used in fraud detection, risk assessment, algorithmic trading, and customer service in the banking and finance sector.
- **Image and Speech Recognition:** AI techniques enable accurate image recognition and object detection. Speech recognition is used in applications like voice assistants, transcription services, and voice-controlled systems.
- **Robotics:** AI-powered robots are used in industries like manufacturing, logistics, and healthcare. They can perform repetitive tasks, assist in assembly lines, and operate in hazardous environments.
- **Gaming:** AI is used in game development to create intelligent and realistic computer-controlled characters, generate procedural content, and optimize game design.
- **Cyber security:** AI aids in identifying and mitigating cyber security threats by analyzing patterns, detecting anomalies, and identifying potential vulnerabilities.
- **Education:** AI applications include intelligent tutoring systems, personalized learning platforms, and automated grading systems that provide immediate feedback to students.
- **Agriculture:** AI helps optimize crop yield, monitor soil health, predict weather patterns, and automate farming processes like irrigation and harvesting.
- **Energy Management:** AI can optimize energy consumption in buildings, predict energy demand, and improve the efficiency of power grids.
- **Social Media Analysis:** AI techniques are used to analyze social media data for sentiment analysis, trend identification, and targeted advertising.
- **Environmental Protection:** AI can be used to monitor and analyze environmental data, track wildlife, detect illegal activities, and optimize resource management.

These are just a few examples of the diverse applications of AI, and the field continues to evolve with new advancements and discoveries.

2. Introduction to Big Data

The architecture of big data [11] typically involves several components and layers that work together to process, store, and analyze large volumes of data. Here is a high-level overview of the architecture:

- **Data Sources:** Big data architectures start with various sources of data, including structured, semi-structured, and unstructured data. These sources can include databases, logs, social media feeds, sensors, and more.
- **Data Ingestion:** The data ingestion layer is responsible for collecting data from the various sources and bringing it into the big data infrastructure. This process involves capturing, validating, and transforming the data for further processing.
- **Data Storage:** Big data architectures employ distributed storage systems capable of handling massive amounts of data. Commonly used storage technologies include Apache Hadoop Distributed File System (HDFS), Apache Cassandra, Amazon S3, and Apache HBase. These storage systems offer scalability, fault tolerance, and high throughput.
- **Data Processing [12]:** Big data processing involves performing computations and transformations on the data to extract valuable insights. There are two main processing approaches:
  a. **Batch Processing:** In batch processing, large volumes of data are processed in periodic batches. Technologies like Apache MapReduce and Apache Spark's batch processing mode are commonly used.
  b. **Stream Processing:** Stream processing handles data in real-time as it arrives. Tools such as Apache Kafka and Apache Flink enable processing data streams in a continuous and near-real-time manner.
- **Data Integration:** Data integration is crucial for combining data from multiple sources and formats. It involves data cleansing, data enrichment, and data transformation to make the data ready for analysis. Tools like Apache Hive and Apache Pig are often used for data integration tasks.
- **Data Analysis:** This layer encompasses various analytics techniques and tools to derive insights from the data. It includes descriptive analytics, predictive analytics, and prescriptive analytics. Technologies like Apache Spark, Apache Hadoop [13], and machine learning frameworks such as Tensor Flow and scikit-learn are commonly used for data analysis.
- **Data Visualization:** Data visualization tools are employed to present the analyzed data in a visually appealing and intuitive manner. Tools like Tableau, Power BI, and Apache Superset help users create interactive dashboards, charts, and graphs.
- **Data Security and Governance:** Big data architectures must address security and governance requirements. This includes ensuring data privacy, access control, compliance with regulations, and data lineage tracking. Technologies like Apache Ranger and Apache Atlas provide security and governance capabilities.
- **Scalability and Fault Tolerance:** Big data architectures are designed to handle the massive scale of data and provide fault tolerance. Technologies like Hadoop's distributed computing framework, fault-tolerant storage systems, and cluster management tools help achieve scalability and fault tolerance.

It's important to note that big data architectures can vary depending on specific use cases and technologies employed. The components mentioned above provide a general framework for designing big data architecture, but the actual implementation can differ based on the organization's requirements and available technologies.

2.1. Process of Big Data

The process of handling big data [2] typically involves several steps, including data acquisition, storage, processing, analysis, and visualization. Here's a high-level overview of the process:

[1] **Data Acquisition:** This step involves gathering data from various sources, which can include structured data from databases, unstructured data from text documents or social media, or semi-structured data from sources like XML or JSON. The data can be collected through sources such as sensors, web scraping, APIs, or data feeds.

[2] **Data Storage:** Once the data is acquired, it needs to be stored in a suitable data storage system. Big data often requires distributed storage systems capable of handling large volumes of data, such as Hadoop Distributed File System (HDFS) [14], Apache Cassandra, or cloud-based storage solutions like Amazon S3 or Google Cloud Storage.

[3] **Data Processing:** In this step, the raw data is transformed and prepared for analysis. It involves data cleaning, filtering, transformation, and integration to ensure consistency and quality. Data processing may involve batch processing, real-time stream processing, or a combination of both, depending on the requirements.

[4] **Data Analysis:** After the data is processed, it is ready for analysis. Various techniques and algorithms can be applied to gain insights from the data. This can include statistical analysis, machine learning, data mining, or predictive modeling. The goal is to extract meaningful patterns, correlations, and trends from the data.

[5] **Data Visualization:** Once the analysis is complete, the results need to be presented in a visual and easily understandable format. Data visualization techniques such as charts, graphs, and dashboards can help in representing the insights derived from big data. Visualization makes it easier for stakeholders to interpret and make informed decisions based on the data analysis.

[6] **Iterative Process:** The process of handling big data is often iterative. As new insights are gained or new data becomes available, the process may need to be repeated or refined to incorporate the new information. This iterative process allows for continuous improvement and adaptation to changing requirements.

It's important to note that the specific tools, technologies, and methodologies used in each step may vary depending on the organization's needs, available resources, and the nature of the data being processed. Additionally, ensuring data privacy, security, and compliance with relevant regulations is crucial throughout the entire big data process.

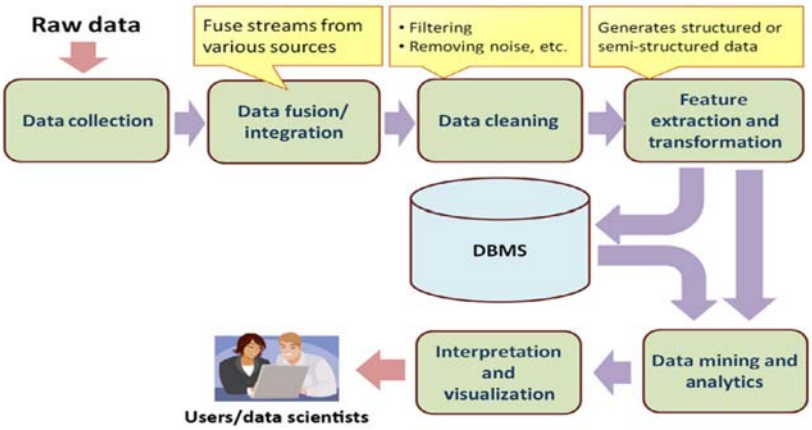

Diagram 3: Understand the Big Data Process

3. Challenges of incorporating big data into AI

Incorporating big data into AI systems brings several challenges that need to be addressed for effective utilization and accurate results. Here are some key challenges associated with incorporating big data into AI:

- ✓ **Data Volume and Scalability:** Big data refers to extremely large and complex datasets that may exceed the processing capacity of traditional systems. Handling and processing such vast amounts of data requires scalable infrastructure and distributed computing techniques to ensure timely and efficient analysis.
- ✓ **Data Variety:** Big data often encompasses a wide variety of data formats, such as structured, unstructured, and semi-structured data. Incorporating diverse data types into AI models requires robust preprocessing and data integration techniques to extract relevant information and convert it into a usable format.
- ✓ **Data Quality and Reliability:** Big data can be noisy, incomplete, or contain errors, which can negatively impact AI models. Ensuring data quality and reliability is crucial to prevent biases and inaccuracies in AI predictions. Data cleaning, preprocessing, and quality assurance techniques are essential to address these challenges.
- ✓ **Data Privacy and Security:** With the increase in data collection and storage, maintaining data privacy and security becomes a significant concern. AI systems must comply with regulations and ethical guidelines to protect sensitive information and ensure data privacy throughout the data lifecycle.
- ✓ **Computational Complexity:** Analyzing large-scale datasets requires significant computational resources. AI algorithms applied to big data often require distributed computing frameworks and parallel processing techniques to handle the computational complexity efficiently.

- ✓ **Model Training and Learning:** Training AI models on big data can be time-consuming and computationally intensive. It may require specialized hardware and software infrastructure to accelerate model training and optimize learning algorithms.
- ✓ **Interpretability and Explain ability:** As AI models trained on big data become more complex; their decision-making process can become opaque and difficult to interpret. Ensuring interpretability and explain ability of AI models is crucial for building trust and understanding their predictions, especially in sensitive domains like healthcare or finance.
- ✓ **Real-Time Processing:** In certain applications, such as fraud detection or real-time recommendations, big data needs to be processed and analyzed in real-time. Incorporating real-time processing capabilities into AI systems requires efficient streaming and processing frameworks to handle high-velocity data streams.
- ✓ **Data Governance and Compliance:** With the use of big data in AI, organizations need to establish robust data governance policies and comply with legal and regulatory frameworks. This involves ensuring data privacy, obtaining necessary consent, and addressing potential biases or discrimination arising from the use of big data.
- ✓ **Cost and Infrastructure:** Incorporating big data into AI systems often requires substantial investments in infrastructure, storage, and computational resources. Organizations must evaluate the costs associated with acquiring and maintaining the required infrastructure and weigh them against the potential benefits.

Addressing these challenges requires a combination of technical expertise, advanced algorithms, scalable infrastructure, and sound data management practices. It is important to carefully consider these challenges and develop strategies to overcome them to unlock the full potential of big data in AI applications.

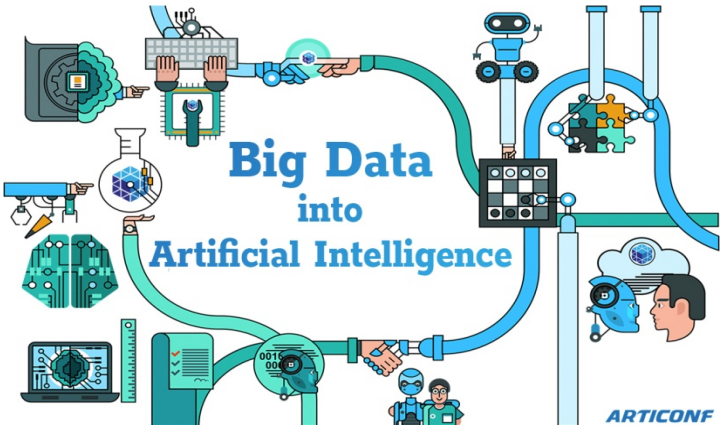

Diagram 4: Challenges of incorporating Big data into AI

Conflicts of Interest

The authors have no conflicts of interest to declare.

Funding

No funding is provided for the preparation of manuscripts.

References

[1] Dr. G. Ravi Kumar and Dr. G. Thippanna An Experimental Concentrate on Impact of Element Assurance in AI approach, Journal Mukt Shabd Journal Volume 11 Issue 10 Pages 339-347.
[2] Big Data Concepts and Techniques in Data Processing by B.Suvarnamukhi, M.Seshashayee, International Journal of Computer Sciences and Engineering
Vol.6(10), Oct 2018, E-ISSN: 2347-2693.
[3] How AI Is Helping Companies Redesign Processes by Thomas H. Devenport, Matthias Holweg, and Dan Jeavons in March 2023.
[4] Krishna Nanda Patel, Sachin Raina and Saurabh Gupta "Artificial Intelligence and its models", Journal of Applied Science and Computations, Vol VII, Issue II, Feb/ 2020.
[5] Artificial Intelligence and Intelligent System by Pat Langley & John E. Laird, January 2006.
[6] Artificial Intelligence Aided Architectural Design by Jan Cudzik, and Kacper Radziszewski, AI FOR DESIGN AND BUILT ENVIRONMENT - Volume 1.
[7] Artificial Intelligence Algorithms by Sreekanth Reddy Kallem, IOSR Journal Of Computer Engineering (IOSRJCE) ISSN: 2278-0661, ISBN: 2278-8727 Volume 6, Issue 3 (Sep-Oct. 2012), PP 01-08.
[8] Introduction to Big data Technology by Bilal Abu-Salih, Pornpit Wongthongtham, Dengya Zhu , Kit Yan Chan , Amit Rudra.
[9] Bohme, F., Wyatt, J. P., & Curry, J. P. (1991). 100 years of data processing: the punchcard century (Vol. 3). US Department of Commerce, Bureau of the Census, Data User Services Division.
[10] https://nitsri.ac.in/Department/Computer%20Science%20%26%20Engineering/BD_11.pdf.

[11] The Hadoop Distributed File System: Architecture and Internals by V. Sajwan, V. Yadav and Dr. M. Haider International Journal of Combined Research & Development (IJCRD) eISSN:2321-225X;pISSN:2321-2241 Volume: 4; Issue: 3; April -2015.
[12] Huang F. (2019) Data Processing. In: Schintler L., McNeely C. (eds) Encyclopedia of Big Data. Springer, Cham
[13] Huang F. (2019) Data Processing. In: Schintler L., McNeely C. (eds) Encyclopedia of Big Data. Springer, Cham

Authors Profile

Dr. G. Thippanna, Professor, Dept. of MCA, working in Ashoka Women's Engineering College, Dupadu, Kurnool, Andhra Pradesh. Received Ph.D. degree in CS & T from Sri Krishnadevaraya University, Anantapuram, Andhra Pradesh, India, in 2016, did Ph. D. in Image processing with entitle of "An Efficient Approach for Image Encryption and Compression using Symmetric Cryptography Techniques". Having 11 years of teaching experience in different institutions. Since 2012 attended many national and international seminars, workshop, FDPs and conferences. And published more than 35 national and international journals in various field especially in image processing, AI and Big Data Technologies. Areas of interest are not in specific, has knowledge in all, but I taught the subjects especially in Big Data technologies, Image Processing, Operating Systems, Computer Networks, Artificial Intelligence, Machine Learning, Data Science, Cryptography Network Security, and Software Engineering, etc…

Dr. D. William Albert, Professor in CSE Department working in Ashoka Women's Engineering College, Dupadu, Kurnool, Andhra Pradesh. Received Ph.D. degree in CSE from Mahatma Gandhi University, Meghalaya, India, in 2017. In 2001, served in the Department of CSE, Sri Krishnadevaraya Engineering College, Gooty, Andhra Pradesh, as an Asst. Professor, and in 2013 became Principal in Shree Vivekananda Institute of Science, Guntakal, Andhra Pradesh, and in 2016 became Principal in Bharath College of Engg. & Tech. for Women, Kadapa, Andhra Pradesh. Since 2015, been submitting National & International Journals. The Ph.D research topic is "A Framework on High Speed Association Rule Mining Using GPGPU-Explorative Studies".Dr. Albert is a Life Member of the Indian Society for Technical Education (ISTE) & Computer Society of India (CSI), and Professional Member in IRED, SDIWC, IAENG.The areas of interest are Software Engineering, Software Project Management, Software Testing Methodologies, Big Data Analytics, Database Management System, Operating Systems and Web Technologies, etc..

Mr. E. Ramachandra, Asst. Professor, Dept. of MCA, working in Ashoka Women's Engineering College, Dupadu, Kurnool, Andhra Pradesh. Received MCA from JNTUA, Anantapuram, India, in 2013. Had 11 years of teaching experience from various institutions. Since 2015 attended two FDP and one Seminar, three workshops.