Chapter 6
Big Data Preprocessing, Techniques, Integration, Transformation, Normalisation, Cleaning, Discretization, and Binning

Pranali Dhawas
https://orcid.org/0009-0003-4276-2310
G.H. Raisoni College of Engineering, Nagpur, India

Abhishek Dhore
MIT School of Computing, MIT ADT University, Pune, India

Dhananjay Bhagat
https://orcid.org/0009-0009-1100-3219
G.H. Raisoni College of Engineering, Nagpur, India

Ritu Dorlikar Pawar
G.H. Raisoni College of Engineering, Nagpur, India

Ashwini Kukade
G.H. Raisoni College of Engineering, Nagpur, India

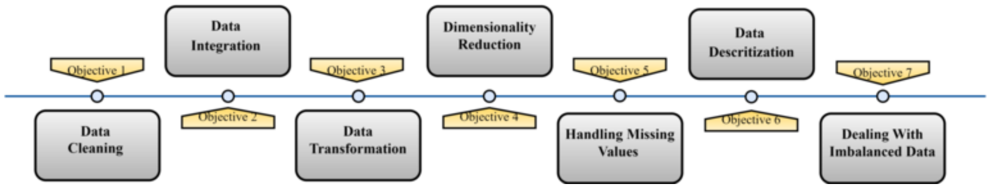
Kamlesh Kalbande
G.H. Raisoni College of Engineering, Nagpur, India

ABSTRACT

"Unleashing the Power of Big Data: Innovative Approaches to Preprocessing for Enhanced Analytics" is a groundbreaking chapter that explores the pivotal role of preprocessing in big data analytics. It introduces diverse techniques to transform raw, unstructured data into a clean, analyzable format, addressing the challenges posed by data volume, velocity, and variety. The chapter emphasizes the significance of preprocessing for accurate outcomes, covers advanced data cleaning, integration, and transformation techniques, and discusses real-time data preprocessing, emerging technologies, and future directions. This chapter is a comprehensive resource for researchers and practitioners, enabling them to enhance data analytics and derive valuable insights from big data.

DOI: 10.4018/979-8-3693-0413-6.ch006

Figure 1. Objectives of big data preprocessing



1. INTRODUCTION TO BIG DATA PREPROCESSING

Big data preprocessing plays a critical role in the data analysis process by converting raw and unprocessed data into a structured and clean format suitable for analysis. As the volume, velocity, and variety of data continue to grow exponentially, preprocessing becomes increasingly vital for extracting valuable insights and knowledge from large datasets.

The process of big data preprocessing involves employing various techniques and operations to enhance data quality, reduce noise and inconsistencies, handle missing values, and prepare the data for subsequent analysis tasks as shown in figure 1. It significantly contributes to improving the efficiency, accuracy, and effectiveness of data analysis (O. Çelik, 2019).

The main objectives of big data preprocessing include:

Data Cleaning: Raw data often contains errors, outliers, duplicates, or inconsistencies. Data cleaning aims to identify and rectify these issues to ensure high data quality. By eliminating noise and irregularities, the resulting clean data provides a reliable foundation for analysis.

Data Integration: Big data originates from diverse sources such as databases, sensors, social media, or IoT devices. Data integration involves combining data from different sources and formats into a unified representation. This step ensures data consistency and compatibility for analysis (Z. Cai-Ming, 2020).

Data Transformation: Data transformation techniques are applied to convert data into a suitable format for analysis. This may involve scaling numerical data, normalizing values, encoding categorical variables, or deriving new features through mathematical or statistical operations. Transformation facilitates data standardization and simplifies subsequent analysis tasks.

Dimensionality Reduction: Dealing with high-dimensional data can pose computational challenges and introduce noise or overfitting problems. Dimensionality reduction techniques help decrease the number of variables or features while preserving crucial information. This simplifies the analysis process and improves computational efficiency (H. S. Obaid, 2019).

Handling Missing Values: Missing data is a common issue in large datasets. Preprocessing techniques include imputing missing values using statistical methods or leveraging imputation algorithms to fill in the gaps. Proper handling of missing data ensures that the analysis is not compromised by incomplete information (T. A. Alghamdi, 2022).

Data Discretization: Discretization involves converting continuous data into categorical or discrete representations. This technique simplifies analysis by reducing the complexity associated with continuous variables. It allows for the application of methods specifically designed for categorical data (P. Gao, 2020).

Dealing with Imbalanced Data: Imbalanced data refers to situations where one class or category is significantly more prevalent than others. Preprocessing techniques address this imbalance by employing