Chapter 8
E-Commerce Data Alchemy:
Behind the Picture of Data Collection and Pre-Processing

Anupam Bonkra

Department of Computer Science and Engineering (CSE), MM Engineering College, Maharishi Markandeshwar (Deemed), India

Pummy Dhiman

Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Mohd Asif Shah

Department of Economics, Kardan University, Kabul, Afghanistan

ABSTRACT

This conversion of raw data into valuable insights is like a complex alchemical process, even more so in the current era where decisions depend on data and are driven by it. It is the work that occurs at a few key stages of data collection and preliminary processing, which is when raw data becomes valuable. This study examines different elements of pre-processing and data collection associated with the complex techniques and steps required during these crucial phases in the data analysis process. These studies seek to establish an overall picture of the fine-grained web of operations and tactics that dictate these crucial phases in data ordering. Expert data analysts and interpreters can capitalize on the tremendous amount of information by meticulously examining and interpreting data. This capability enables them to transmute raw facts into insightful knowledge, fostering innovative thinking and informed decision-making.

DOI: 10.4018/979-8-3693-5718-7.ch008

1. INTRODUCTION

Converting raw data into meaningful human insights is like a high-level synthetic process played out in your brain— but happening in real-time today within the realm of data-driven decision-making, (Rahman, 2023). This approach delves into the initial stages of data science: where raw, unstructured, and unprocessed data is transformed into a valuable asset. Our aim through this introduction is to provide a holistic understanding of the strategies involved in these critical phases of data analysis. An overview on the complexity surrounding data collection and pre-processing/cleaning will be discussed shortly. The methodology for acquiring data acts as the bedrock for any subsequent analysis; it entails gathering unstructured text repositories, structured relational databases, sensor-generated data streams— or any other forms of raw data, no matter how diverse they may be. However, the quality of the data, its capacity to be used, and its representativeness are what make data gathering worthwhile, (McDermott *et al.,* 2024). It is not the quantity of data collected that is the most important factor in determining its value. As a consequence of this, stringent procedures are used for the purpose of systematically verifying robustness and variation in order to develop good analyses and insightful discoveries based on data that is given as appropriate evidences, (Rahman, 2023). In the process of processing raw data as if it were pre-existing ore and putting it through a crucible of refining flames to transform it into genuine, actionable value (while also considering your own interests), the alchemical processes that are taking place should be seen as focusing and rally points. In this step, a range of activities are included, such as methods for feature engineering, which are used to find important data patterns and features; noise reduction strategies, which are used to remove misbalances and errors in the produced data sample; and imputation techniques, which may minimize the number of missing values, (Yalamati & Batchu, 2024). In the realm of big data, which is characterized by datasets that are both bigger and more sophisticated, it is essential to have efficient pre-processing procedures in order to have successful modelling. By utilizing strategies like dimensionality decrease, exception identification, and information pressure, experts had the option to deal with tremendous informational indexes in a short measure of time and get huge experiences from the mind-boggling amount of information that was available. Our goal is to dive into the perplexing field of information speculative chemistry and uncover the lost fortunes of pre-handling and information assortment. These subtle landscapes highlight a mind boggling organization of procedures and approaches and act as guards to these significant stages in the information examination venture. For analysts and data scientists to unlock the dormant value in their datasets, it is necessary to have an understanding of the intricate processes that go into data acquisition and pre-processing, (Rahman, 2023).. However, this symbiotic relationship between