Chapter 3
Big Data Analytics Lifecycle

Smrity Prasad
https://orcid.org/0000-0002-8057-3612
Christ University, India

Kashvi Prawal
Penn State University, USA

ABSTRACT

Big data analysis is the process of looking through and gleaning important insights from enormous, intricate datasets that are too diverse and massive to be processed via conventional data processing techniques. To find patterns, trends, correlations, and other important information entails gathering, storing, managing, and analyzing massive amounts of data. Datasets that exhibit the three Vs—volume, velocity, and variety—are referred to as "big data." The vast amount of data produced from numerous sources, including social media, sensors, devices, transactions, and more, is referred to as volume. The rate at which data is generated and must be processed in real-time or very close to real-time is referred to as velocity. Data that is different in its sorts and formats, such as structured, semi-structured, and unstructured data, is referred to as being varied.

INTRODUCTION

Big data arose in recent years to fulfill the demands and challenges of expanding data volumes. Big data refers to the process of managing massive amounts of data from many sources such as DBMS, log files, social media posting, and sensor data (Bajaj et al. 2014). When we hear the term "big data," we immediately think of the massive amounts of data that must be stored and processed. Indeed, a large volume of data is a big data type feature that exceeds Exabyte (1018), necessitating unique storage solutions, high performance data processing, and particular analytics capacity (Kaisler et al., 2013). Big data is a collection of complex datasets (text, numbers, photos, and videos) in massive volumes that exceed the capabilities of typical database management systems (Govindarajan et al., 2014).

Big data, in particular, has three primary characteristics: volume, velocity, and variety. Aside from the three Vs, other big data traits included value and complexity (Kaisler et al., 2013; Katal et al., 2013).

DOI: 10.4018/979-8-3693-0413-6.ch003

The volume attribute denotes the amount of data. In general, big data has a vast volume of data that is beyond the capacity of typical storage systems. According to (Bajaj et al.,2014), 90 percent of the world's current data was created in the last two years, with an average of 2.5 quintillions of data bytes created everyday. The velocity aspect of big data relates to the rate at which data is generated and processed (Bajaj et al.,2014). Currently, data and information are generated and processed at a high pace, resulting in a massive amount of knowledge being contributed to the knowledge base; this velocity rate of big data necessitates more processing power than older systems. Furthermore, the term velocity alludes to the rapid movement of data between data storage locations via networks (Bajaj et al., 2014).

Variety is another important aspect of huge data. The term "variety" in big data refers to the various resources that generate data in various formats and types (Bajaj et al. 2014; Govindarajan et al. 2014; Kaisler et al., 2013; Katal et al., 2013). Digital photographs and videos, social media, sensor data, healthcare data records, text, log files, tweets, and purchase transaction records are all examples of data resources. In other words, big data is made up of several data forms, including structured, unstructured, and semi-structured data.

Value and complexity are two further big data properties (Kaisler et al., 2013). The value attribute in big data refers to the usefulness of information (knowledge) that may be derived from processing and analyzing big data. This newly produced information is beneficial and supportive of decision-making (Katal et al., 2013). The complexity attribute refers to the complexity of relationships and links in a large data structure. In this regard, we may understand how complex it is when only a few changes occur in enormous amounts of data, resulting in a significant number of modifications (Katal et al., 2013).

The Big Data process involves a number of processes, beginning with data collection and ending with decision-making. Researchers agree on a few key factors regarding the steps. As an illustration, (Bizer et al.,2012) list six (6) steps: data collection, storage, search, sharing, analysis, and visualization. (Marx, 2013) suggests five (5) phases to tackle a problem: problem definition, data searching, data transformation, data entity resolution, and query response/problem resolution. (Chen and Liu, 2014), in contrast, solely employ the three steps of data handling, data processing, and data movement. Building competences and capacity for data management is necessary for an efficient big data chain .The outcomes are influenced by the capacities of each business involved in the big data chain of information. Therefore, depending on the capacities of each entity, the capacity of organizations and enterprises to gather, prepare, and analyze big data may vary.

After analyzing number of authors, Following are the nine steps that make up the Big Data analytics lifecycle. Life cycle of big data analytics is shown in Figure 1.

1. Business Requirement Identification and Evaluation
2. Data Identification
3. Data Acquisition & Filtering
4. Data Extraction
5. Data Validation & Cleansing
6. Data Aggregation & Representation
7. Data Analysis
8. Data Visualization
9. Utilization of Analysis Result