Big Data

Sanjay Reddy Muthyala

May 3, 2022

1 Introduction

Big Data is an extensive collection of data that continues to grow dramatically over time. Big data is similar to regular data, but it is much larger. Big data collects organized, semistructured, and unstructured data that may be mined for information and used in machine learning, predictive modeling, and other advanced analytics initiatives.

2 Importance

Companies employ big data in real-time systems to enhance operations, provide better customer service, generate targeted marketing campaigns, and take other activities to boost revenue and profitability. As a result, businesses that properly use it have a potential competitive advantage over those that don't since they can make more informed and faster business decisions. Big data, for example, provides firms with important customer insights that they can utilize to improve their marketing, advertising, and promotions to boost customer engagement and conversion rates. In addition, consumer and corporate purchasers' developing preferences can be assessed using both historical and real-time data, allowing organizations to become more responsive to their wants and needs. Medical researchers and doctors use big data to uncover disease indicators and risk factors and diagnose illnesses and medical problems in patients. Furthermore, data from electronic health records, social media sites, the internet, and other sources are combined to provide healthcare organizations and government agencies with up-to-date information on infectious disease threats and outbreaks. Big data assists oil and gas businesses in identifying new drilling locations and monitor pipeline operations, while utilities use it to track power networks. Financial services firms use big data systems for risk management and real-time market data analysis. Big data is used by manufacturers and transportation businesses to manage supply chains and optimize delivery routes.

3 Three V's

The term "Big Data" may be traced back to the mid-1990s, when it was initially utilized to handle and analyze massive datasets by John Mashey, a retired former Chief Scientist at Silicon Graphics (Diebold, 2012). Doug Laney described three characteristics that constituted big data in 2001.

- Volume: The storage capacity necessary to record and store data is referred to as volume in the context of Big Data. As it is generally stated, Big Data requires terabytes (240 bytes) or petabytes (250 bytes) of storage capacity, significantly more than a standard desktop computer can supply. The data is frequently kept in the cloud across multiple servers and regions. Human-mediated forms, such as those used to create administrative records from immigration and unemployment registration, may generate a steady stream of new records from a few sites, such as a few entries pointing to an unemployment office, resulting in much lower volumes than automated systems. Similarly, while each sensor record is often tiny in file size, imaging data such as streaming video, pictures, and satellite images are frequently enormous, so even small numbers of records quickly add up to significant storage requirements. Even though the amount per record is often minimal, many devices generating data results in massive storage volumes.

- Velocity is the rate at which these massive amounts of data are generated, collected, and processed is called velocity. Big Data is known for its speed, which is regarded as one of its most essential characteristics. Rather than sampling data on a one-time basis or with a vast temporal gap between samples, Big Data is produced on a much more regular basis. However, when we looked at our datasets, we discovered two types of velocity when it comes to Big Data frequency of generation, handling, recording, and publishing. In terms of generating frequency, data can be generated in real-time at any moment. For example, at the point of use, such as clickstream data being generated in real-time but only while a user clicks through websites, or an immigration system recording only when someone is scanning their documents.

- Variety refers to the different types of data that no longer have only structured data that can be appropriately displayed in a data table like name, phone number, and ID. However, current data is unstructured primarily: images, audio, social media updates. This seems to be the weakest attribute of all the characteristics attributed to Big Data. Indeed, small data are also highly heterogeneous, especially datasets common to humanities and social sciences, where the handling and analyzing of qualitative data like text and image is normal. We suspect that this characteristic was attributed to Big Data because those scientists who first coined the term we are used to handling structured data exclusively but were starting to encounter semi-structured and unstructured data as new data generation and collection systems were deployed.

4 References

- Rob Kitchin1 and Gavin McArdle2: What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets

- Big Data By Bridget Botelho, Editorial Director, NewsStephen J. Bigelow, Senior Technology Editor.