

The Four V's of Big Data

Big data environment has these four basic characteristics – volume, variety, velocity and veracity.

The volume of data in the last two decades has grown exponentially, (see lesson 1 – value of data). However, the cost and efficiency of storing this huge amount of data reduced and increased significantly over the last two decades.

The velocity refers to how efficiently we transmit the data through the internet and other forms of wireless or wired networks. Again, within the last two decades, this number and type have increased many folds with the introduction of fibre optics and laser technologies.

The last two decades also witnessed a huge transformation in terms of data types or variety. Previously it was mostly structured textual data but with the advent of digital cameras and social media, there is a tsunami of audiovisual and unstructured data (free text). Interoperability within different data formats remains a big challenge for the next generation of data scientists.

The veracity or uncertainty of big data refers to both data quality and availability. Although 2.5 quintillion bytes of data are generated every day and a big part of those data are open to public use, businesses are often unsure about the quality of the data, that is how the data was collected? What is the source of the data? etc. Such uncertainty reduces the use of the data for further analysis and thereby the value of the data. Availability of data at the right time and place is another factor. This is more prevalent in healthcare as most of the healthcare data are personal data and therefore privacy of the data needs to be ensured before any use of the data.