A "Small" Definition of Big Data

The term 'big data' seems to be popping up everywhere these days. And there seems to be as many uses of this term as there are contexts in which you find it: 'big data' is often used to refer to any dataset that is difficult to manage using traditional database systems; it is also used as a catch-all term for any collection of data that is too large to process on a single server; yet others use the term to simply mean "a lot of data"; sometimes it turns out it doesn't even have to be large. So what exactly is big data?

A precise specification of 'big' is elusive. What is considered big for one organization may be small for another. What is large-scale today will likely seem small-scale in the near future; petabyte is the new terabyte. Thus, size alone cannot specify big data. The complexity of the data is an important factor that must also be considered.

Most now agree with the characterization of big data using the 3 V's coined by Doug Laney of Gartner:

· Volume: This refers to the vast amounts of data that is generated every second/minute/hour/day in our digitized world.

· Velocity: This refers to the speed at which data is being generated and the pace at which data moves from one point to the next.

· Variety: This refers to the ever-increasing different forms that data can come in, e.g., text, images, voice, geospatial.

A fourth V is now also sometimes added:

· Veracity: This refers to the quality of the data, which can vary greatly.

There are many other V's that gets added to these depending on the context. For our specialization, we will add:

· Valence: This refers to how big data can bond with each other, forming connections between otherwise disparate datasets.

The above V's are the dimensions that characterize big data, and also embody its challenges: We have huge amounts of data, in different formats and varying quality, that must be processed quickly.

It is important to note that the goal of processing big data is to gain insight to support decision-making. It is not sufficient to just be able to capture and store the data. The point of collecting and processing volumes of complex data is to understand trends, uncover hidden patterns, detect anomalies, etc. so that you have a better understanding of the problem being analyzed and can make more informed, data-driven decisions. In fact, many consider value as the sixth V of big data:

· Value: Processing big data must bring about value from insights gained.

To address the challenges of big data, innovative technologies are needed. Parallel, distributed computing paradigms, scalable machine learning algorithms, and real-time querying are key to analysis of big data. Distributed file systems, computing clusters, cloud computing, and data stores supporting data variety and agility are also necessary to provide the infrastructure for processing of big data. Workflows provide an intuitive, reusable, scalable and reproducible way to process big data to gain verifiable value from it in and enable application of same methods to different datasets.

With all the data generated from social media, smart sensors, satellites, surveillance cameras, the Internet, and countless other devices, big data is all around us. The endeavor to make sense out of that data brings about exciting opportunities indeed!

Source: http://words.sdsc.edu/words-data-science/big-data

✓ Complete