Stanford eCorner

What is Big Data?

13-11-2013

URL: https://stvp.stanford.edu/clips/what-is-big-data

Mike Olson, co-founder and chief strategy office of Cloudera, shares multiple definitions of “big data.” While the accepted industry definitions think of big data in terms of volume, velocity, and variety, Olson describes big data as being, “any amount of data that doesn’t fit where you want it.”

Transcript

I'm going to give you sort of the broadly accepted view and then I'm going to actually tell you what I think.. So it's possible to get your hands on lots of data and it's possible for you to get your hands on lots of kinds of data, and it's possible for you to get your hands on a fire hose of data these days, right, any one of those can happen to you.. You can get transactional data and tweets and weather models and that complexity, my goodness it's a bunch of different stuff.. And you can get literally petabytes easily, it's hard not to get terabytes these days, that used to be a scary number and now we're saying oh yeah it's only 100 terabytes, it's hardly anything.. And the volume, the velocity rather at which it arrives can be overwhelming.. Any one of those by themselves could feel like big data to you.. If you've got two or more of those problems it could be absolutely overwhelming.. So volume, variety, velocity, those are the memes that the industry has generally embraced.. My point of view is, if there is data you want to work with and it doesn't fit where you want to put it, it's big data.. And that might be, because it's complicated and you got lots of different kinds..

It might be because you want to ask way more complicated questions of that data than were previously possible.. So there is a very common until recently research technique now widely used in the industry called machine learning and actually some of that work has been done here at Stanford.. I would say probably better work had been done at Berkeley, but machine learning requires good systems for scaling out very broadly and machine learning is an unbelievably powerful technique for deducing facts from data that aren't obvious using other techniques.. You can let the data tell you what it's about and what's happening.. And machine learning really, really likes Hadoop.. It really likes these scale out architectures.. So if you wanted to do that kind of analysis, if you want to store that variety, if you want to land a petabyte somewhere and you can't afford to spend $40,000 a terabyte to do it, turns out this platform is pretty good and big data is that general meaning...