# Defining Big Data

Bob Gourley | August 24, 2012


Enterprise IT professionals, including public CIOs, have long recognized the power of data, and the exciting new sense-making capabilities around big data approaches have generated a great deal of buzz and excitement. If history is a guide, however, we are about to see that term lose much of its meaning. Here is what I mean:

Do you remember service-oriented architecture (SOA)? This concept led to tremendous new capabilities and efficient, mission-focused designs. Enterprises established architectures in which application interfaces, logic and data were separated and smartly reusable. After the term went mainstream, every company in the IT ecosystem grabbed onto it and began to use the acronym SOA to mean anything they wanted it to. Although it's still a useful construct for IT professionals, when it comes to interacting with industry, the term has now lost much of its meaning.

Then there's cloud computing. When enterprise IT professionals use that term among themselves, there is huge value in the concept. It conveys a great deal of meaning regarding a need to change business processes to take maximum advantage of modern IT and new offerings. Now, however, most IT vendors describe what they do as cloud

computing. When it comes to interacting with industry, that term, like SOA, has lost much of its meaning.

Now what about big data? Today it remains a very helpful term. Practitioners, including IT architects, systems engineers, CIOs, CTOs and data scientists, all use this term in dialog over ways to improve sense-making over data. The term remains a useful way of introducing others, including non-technologists, to new approaches like the Apache Hadoop framework. We have a continuing need to discuss these topics, and the term “big data” will likely be with us for quite a while.

But just like SOA and cloud computing, big data is now a hot topic among the vendor community. All indications show that most IT vendors are aware of the exciting dialog under way on this term. All have either already shifted their marketing strategy to include this topic — or they soon will. Odds are that most every firm in the IT industry will soon be proclaiming itself to be a big data company.

I’ve already seen plenty of evidence that this rebranding is under way. I’ve heard makers of network switches and routers assert that they are big data companies because they move large amounts of data. I have met with mapping companies that want to be called big data companies because they plot data. I know of an old-school storage company that wants to be known as a big data company because it stores lots of information. A great information integration company I know and love has told me it’s the big data solution of choice since it integrates data. The leading chip-maker is about to kick off a big data campaign, because it takes processors to process big data.

And in every case, the firms are creating their own definitions of what big data is. History is going to repeat itself here. Very soon, every vendor you deal with will want to get you to use its definition of big data.

So what should public-sector technologists do in an environment like this?

I recommend doing what enterprise technologists do best: Focus on your mission needs; don’t let anyone convince you to conform to their concepts of how those needs should be met.

And when it comes to definitions, you should be prepared to articulate one that best meets your organization’s needs. As a starting point, I recommend the definition at Wikipedia.org, since this community-edited site captures the input of many. Wikipedia’s definition is this: “Big Data implies the need for a strategy for dealing with large quantities of data. The term is also used to describe the new platform of tools required to successfully handle sense-making over large quantities of data, as in the Apache Hadoop Big Data Platform.”

I like this definition because it focuses on sense-making over data, which is why we have the data to begin with. I also like the reference to Apache Hadoop, since every big data solution I know of uses this framework. Hadoop is usually key to big data, but other important capabilities in this framework include HDFS, HBase, Hive, Cassandra and Mahout.

If you select a definition that doesn't key in on sense-making over data, then you automatically open yourself up to letting every maker of any IT capability say it is a big data company. And if you don't mention the Apache Hadoop framework in your definition, you open yourself up to allowing every maker of legacy software to say it is a big data company even though it has the same old approach. There's something new about big data designs, and that is the distributed processing of large data sets over clusters of computers enabled by the Hadoop framework.

Whatever definition you decide to use, I would recommend you dive deep into learning the capabilities of the Apache Hadoop software library. This framework enables distributed parallel processing of huge amounts of data across inexpensive, commodity servers — and no vendor should bring you a big data solution unless it has leveraged the powerful capabilities of this framework.

Big data and how the community uses the term is a topic in need of more discussion, and my hope is that technologists from across the public sector, at local, state and federal levels, have a greater dialog on what that term means to public-sector missions. Discussing this topic could prove to be very positive for organizational missions and will help the IT vendor community better understand public-sector needs.

*Bob Gourley is the editor of CTOvision.com and is the founder and Chief Technology Officer (CTO) of Crucial Point LLC, a technology research and advisory firm.*

http://www.govtech.com/pcio/Defining-Big-Data.html