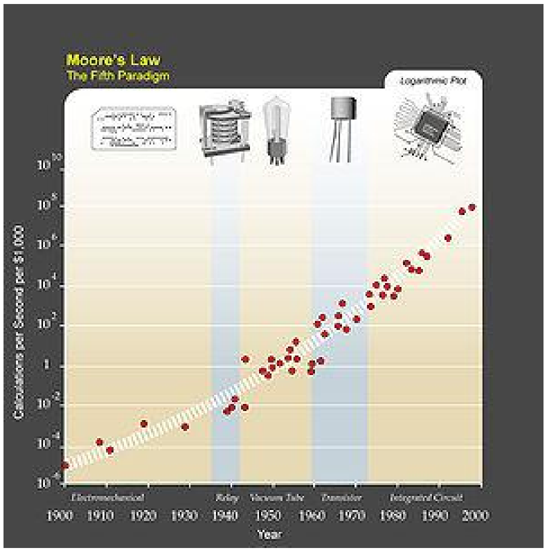
5,522 views | Dec 12, 2013, 11:40pm

How To Create A Moore's Law For Data

Dan Woods Former Contributor ⓘ
Enterprise & Cloud

This article is more than 2 years old.



Moore's Law, The Fifth Paradigm. (Photo credit: Wikipedia)

We are often reminded in press and analyst reports th more data has been created in the last year than in all previous years combined. Such articles often are written in a giddy tone based on the unstated assumption that more data will mean more value, mo benefit to us all.

At first glance, this seems like a reasonable proposition. More of something (money, time, food) often means that more benefit can be obtained. I suspect the authors of such articles have Moore's La in mind, which, in its popular understanding predicts the ever increasing power of computers.

Today In: Tech

Because we have an intuitive understanding of data, often we don't ask important questions. W know what data is, how it can be stored, how to move it, how to analyze it, and in a sense, we think we understand the nature of data.

But a closer look at the world of data shows that there is no Moore's Law in effect. More data just means more data. In many cases data is a liability. More data means more costs for storage, for governance and having too much unorganized data may make it more difficult to find what you need. In other words more data can mean less value.

A closer look at Moore’s Law shows that it only works because there is an existing and well developed, multi-layered ecosystem that leads from more of something (transistor density) to more benefit.

If we draw an analogy from mechanisms of Moore’s Law to the world of data, we can see just what we need to do to make more data mean more value. It is time to stop paying attention to th growing volume of data and start working on the mechanisms that will allow it to help create a better world.

Why Moore’s Law Works

Moore’s Law states that the number of transistors on integrated circuits doubles every 2 years. But if you asked most people what Moore’s Law means, some would say that Moore’s Law predicts that computers will get faster and cheaper at a rapid rate, a statement that turns out to b true. Others would say that Moore’s Law means the computers become more powerful. Also tru It is important to understand why.

There are two stages in the trip from more transistors on chip to more value for us all. The first the hardware industry, which uses the increased number of transistors to create more powerful chips, which are used to create more powerful computers. The second is the software industry, which uses more powerful computers to create more powerful applications.

Because of the hardware and software industries, Moore’s Law actually works the way most people think it does.

How a Data Stack and Data Economy Could Lead to a Moore’s Law for Data

To make a Moore’s Law for data, we also need two layers, a data stack and a data economy lay If both of these layers were as mature as the hardware and software industries, more data would mean more value. But these layers are just getting off the ground. I suspect most people looking to take advantage of the glut of data will benefit from thinking about how to create their own da stack and how to put it to work in the context of a data economy.

Gil Elbaz, co-founder and CEO of Factual, an innovator in creating high quality data sets about products and places from a diverse set of sources, defined the concept of the Data Stack Layer i this article: “The Data Stack: A Structured Approach”. The data stack defines all the capabilitie needed to connect to data, collect it, clean it, and join it together into a form that makes sense. Then the data stack describes capabilities for delivering that data to those who use it in multiple ways, through subscriptions or through APIs.

The Data Stack differs from the way most current data processing systems work for a variety of reasons.

- Much of today's data processing and analysis is implicitly bound by "not-invented-here" thinking. (See "Do You Suffer from the Data Not Invented Here Syndrome".) Business intelligence and data warehouse systems usually only collect data that is inside the four walls of a business. A data stack is like an extended supply chain that is gathering data a multitude of sources, both internal and external.

- Often a single data source and a data set are equated. In a data stack, a data set is assembled from fragments coming from a wide and heterogeneous set of sources.

- Most of the time, companies assemble data sets using a single processing pipeline. A da stack may have multiple pipelines for different purposes, one for example to assemble tl data set as fast as possible and another aimed at quality that takes more time to check fo errors with advanced algorithms.

- In most data processing systems, data the describes where data comes from, that is the provenance of the data, doesn't exist or doesn't travel along with the data. In a data stac provenance is vital because there are so many sources being combined.

- Data stacks assume that data must be curated and maintained, in other words data is botl an asset and a liability. Often, the fact that a data set is actually a liability is ignored in much data processing. In data stacks, the responsibility for curation is addressed as part the initial design.

- While data warehouses may hive off portions of data into data marts, data stacks have multiple distribution methods such as APIs that are purpose-built to support various use In addition, the subscribers to data created by a data stack are often different organizations, while data marts are usually internal.

If we imagine a world in which the plumbing for data stacks was well-understood and widely implemented, then more data would quickly lead to more value. In my view, the paradigm of th data warehouse must be replaced by the idea of a data supply chain. (See “Why Building a Distributed Supply Chain is More Important than Big Data”.) But that is only the first step.

The second step is to motivate all of the participants who have data, who could assemble clean, and curate data, who could build new products from data, and who could use the data to work together. When everyone is motivated to work together a self-supporting data economy is created. Remember, a data economy could exists inside the four walls of a company, but, it is likely that the most powerful data economies will span many companies, like supply chains. Th is exactly what Factual is building.

Right now, each data economy is being crafted in a custom way. The large data vendors such as Equifax, Experian, Acxiom Corporation, Harte-Hanks, and infoUSA are all data economies bui in a traditional way, primarily according to the internal data economy. But eventually, common mechanisms and roles will emerge. Here is my first guess at the roles needed to create a data economy that works more like a inter-company supply chain:

- Providers, those who have data that may be able to help create a valuable data set.
- Curators, those who collect data from providers, use technology and other means to create valuable datasets, and then provide it to developers and consumers.
- Developers, those who use the data from curators and providers to create or enhance products.
- Consumers, those who make use of the data directly or through products.

Understanding the principles of a data economy is important because it can focus efforts on wh needs to be done to make data accessible and catalyze the creation of such an economy. Companies should be thinking of their natural roles. Are you data collectors, distributors, curators, application developers? Companies should be thinking of what advantages they have i creating the missing technologies. It is likely that internal resources in a company could be bett exploited if these roles were defined more clearly and explicitly supported.

The trick is jumpstarting such economies so that an ever expanding amount of valuable data set are created. But there are a variety of barriers to getting data economies started that Elbaz has explained to me during some research interviews last year.

One huge barrier Elbaz points out is the assumption of altruism that seems to motivate many efforts at sharing data, especially projects that fall under the open data umbrella. When organizations consider sharing data, they often incorrectly assume that they must just give the data away and hope that some benefit accrues to someone. Government initiative have resulted many victories using this model. But organizations are sitting on massive troves of data that won't be shared unless there is a fair trade.

Another barrier Elbaz explained is a limited understanding of standard models and commercial terms that can motivate sharing of data. Think of sharing like putting money in a bank. You are sharing your money because it is safer in the bank, because you get interest in for your money, because you can better use your money when it is in a bank. The bank uses your money to mak loans and benefit others, but you benefit as well. Sharing data with some sort of curating organization at the center of a data economy should be motivated by similar rewards.

The good news is that the number of people who understand the nature of data economies are growing. The Green Button initiative in which consumers can get access to their energy usage data is a fine example of the power of releasing data. Third party services can use this data to help make people aware of how they can better optimize energy use. But this is a data economy that is encouraged by regulators and falls short of a self-sustaining data economy.

At Factual, Elbaz and his team are creating data economies around Global Places and Global Products datasets. Dozens of organizations and thousands of individuals are participating by playing all the roles mentioned above.

These are still early days for any sort of Moore's Law for Data. You can see Google, Facebook, and Twitter as data economies, perhaps ones in which the distribution of value is too much in favor of the curator. But there is so much room for more creative structures and so much data t be put to use. The potential to create value with new forms of data economies is vast and under-appreciated.

In my view, the excitement about the growth in the volume of data is misplaced. Instead, peopl should be energized by the opportunities to create data economies. That the barriers listed abov are significant represents barriers to entry of the sort that should excited entrepreneurs or companies seeking a defensible competitive edge.

As the technology of the data stack improves and becomes more productized, as the nature of how data economies is more widely understood and more and more successful examples appear I firmly believe we will have a foundation for a Moore's Law for data. When we get there, mor data will indeed more value and the world will be a much better place.

Follow Dan Woods on Twitter:

Follow @danwoodscito

Dan Woods is CTO and editor of CITO Research, a publication that seeks to advance the craft technology leadership. For more stories like this one visit www.CITOResearch.com. Dan has performed research for Factual and many other data-related companies.

Dan Woods

My mission: Find technology for Early Adopters. Follow me: on Twitter @danwoodsearly on LinkedIn @ www.linkedin.com/in/danwoodsearly/ on myBlog @ https://earlyadopter.co... **Read More**