Stanford eCorner

Evolution in Data Storage and Management

Mike Olson, *Cloudera*

November 13, 2013

Video URL: http://ecorner.stanford.edu/videos/3224/Evolution-in-Data-Storage-and-Management

Cloudera Co-Founder Mike Olson quickly steps through how data management techniques have changed since the late 1990's, and how these changes have opened wide, new opportunities for the future.

Transcript

I'm a old guard relational database guy. I built and sold relational database products and then I built and sold companies that built and sold relational database products for my entire career. So I was very deeply steeped in how those systems work and what they were good at, what they were not so good at. Back in the day, if you wanted to manage data, what you did was you called up Sun Microsystem and you had them ship you the very biggest box they had, right and you wrote them a cheque for that. And if you had a little bit of money leftover, you would call up Oracle and you would get some database software to run on that one big computer and that was your data temple, that's where everything went. Big centralized servers were how we built systems back then. Then you could connect to that from lots of places, but all your data had to be in that one big box. And the way we built systems assumed one big box. It was just a single computer that stuff ran on. When Google wanted to index the entire Internet, it turned out there was no box big enough.

You couldn't buy a single computer that would fit the entire web even in 2000. Google didn't realize what we knew in the database industry and that was - it was impossible to build massive scale out data management infrastructure. So they just went ahead and did it, and you Stanford guys, just piss me off by the way. Google invented a scale out platform that would do that and the way that they did it was instead of one big Sun Microsystems box, a whole bunch of little computers that were ganged together. All of them have local disk, all of them have local processing, you get a bunch of data and you bust it into pieces, just peanut butter, spread it around all those servers. You know what, you're going to lose some of those servers, because they're cheap and unreliable, so just store multiple copies and let the software account for those failures. So you can store the data really cheaply; not just store it, all those computers have a lot of processors on them, so you have got a bunch of computer power distributed among all your data. So if you want to ask a question of a whole bunch of data you send that question out to all those computers and they all look at their own little piece of data, reason about it, and produce an answer. And that's - honestly, you guys, it's a miracle. You can run a query on 100 terabytes of data and get an answer back in minutes.

And you can buy 10 times more computers, and you can ask that same question of a petabyte of data and you get an answer back in the same number of minutes. This is unheard of; this had never happened before in the industry. So that platform was really transformative and Google was able to use it to index the web, to observe user behavior. They want to continually improve their search results, so they watch the links you click on and automatically adjust the way they return results to others running similar searches. That processing power was really transformative for them. I had left Oracle, I was looking for something new to work on, I recognized that this was just a vastly different architecture than we've been building for decades. And simultaneously two things had happened in the industry. One was you could buy cheap computers from a lot of vendors. So this commodity hardware architecture totally had happened. The way we build data centers now is not one big computer; it's just racks and racks and racks of commodity boxes.

But another important trend happened and that was all of you guys started carrying around cell phones. All of you guys started using Twitter and Facebook and the volume of data basically machine generated data, telemetry from where you are, the transponder in your car that talks to tollbooths and road sensors, data was being generated at machine scale and that was new. So three things happened really simultaneously: interesting new compute and storage platform invented by Google that ran on newly, cheaply available commodity hardware that you could get from a lot of vendors, at a time when data volume, data variety, data velocity were all exploding. That three-plex of things really created the perfect wave. Jeff and Amr and I and Christophe, the four of us who were looking, we all saw all of those trends at the same time, we all recognized that they were going to be a big deal, not for consumer internet, they'd been successful there, but for banks and insurance companies and hospitals. The conviction that big data would happen broadly was what really brought us together and what convinced us that there was a big opportunity to commercialize this software. And I will say, touch wood we were the first. I think we were far seeing, I think we've executed pretty well. But the market very deeply believes this now; there is lots of investment, lots of activity. The opportunity has grown just tremendously in the last several years.