Stanford eCorner

Opportunities Abound in the Big Data Space [Entire Talk]

13-11-2013

URL: https://stvp.stanford.edu/blog/videos/opportunities-abound-in-the-big-data-space-entire-talk

Cloudera Co-Founder Mike Olson shares his insights on the present landscape and possible future of big data and the data management industry. In conversation with Ping Li of Accel Partners, Olson also discusses the advantages of building a business on top of open source technologies and the many surprising benefits of competition.

Transcript

Good to see you, Ping.. Yes.. So I guess the first question is what's it like to be at Stanford being a Cal guy? I spend actually a fair bit of time on campus here.. We are headquartered very nearby just across Page Mill Road right at the corner of Page Mill and El Camino and we get a lot of opportunities to come and speak on campus and to engage with the students and we have got a very successful internship program that recruits from here and I make it a point to donate a little bit extra money and time to my alma mater as a result of my engagement down here.. So it is good for everybody.. All right.. Well hopefully the audience will take it easy on you today.. So, I'm sure a lot of folks have read about the founding of Cloudera and how it got started.. But I'd love to hear it from your perspective and not the bloggers, how do you - how would you describe the early days, pre-funding, pre-everything? Yes, you bet.. So for those of you who don't know what Cloudera does, is we commercialize some software that was originally invented at Google..

And the software is called Hadoop and it's for working with lots of data, big data in ways that you never could before.. In the early part of 2008 I had sold my previous company to Oracle and I did two years of lock-up at Oracle after that acquisition and it was a good experience, but when the two years ended I left, I was looking for the next thing to do.. I began looking around for interesting opportunities in the industry and I came across Hadoop.. And I have heard about it before, but not paid much attention.. When I dove in with some of the big users of the software, Facebook, Yahoo and others and understood what they were doing, I got super excited and I wanted to start a company around it.. So I did what you do, right.. I walked around and talked to everybody I knew.. And in the process of doing that I found three other guys who had decided that there was an opportunity to create a company around Hadoop and all four of us had an idea of what that company was going to go do, so sort of early summer of 2008 there were four different people looking to create a company on a technology no one's ever heard of in pursuit of a zero dollar market.. Did not .... Sounds great..

.... yes, recipe for success, right.. Two of the people who were looking at doing that were actually being at the time incubated at Accel.. Ping had found these guys at Facebook and at Yahoo respectively.. The four of us just by networking around and looking for others working on this found one another.. We spent really the summer trying to decide if we could collapse four cap tables into a single cap table and create one company.. And we successfully did that and let me say, had we not I think it would have been a disaster.. It would have been a lack of focus and it would have been very distracting for all of us.. By the end of the summer, we all believed that we could work together, we spent a little bit of time in the very early fall, putting together a strategy and thinking about how we would raise money, Accel had incubated Jeff and Amr as I said in order to encourage them to go and do something interesting, was interested in funding and Ping was the partner that had driven that.. That's how I got introduced to him..

I knew Accel quite well.. But given that, we had the four people thinking about starting a company and the one investor in the Valley who was interested in doing something to do, I figured let's just take all this stuff off the market.. So we wound up raising our series A round from Accel.. Ping joined the board with the funding in October 2008.. Just one side story, I don't know well - how well you guys remember it.. Those were dark days, right? Like a week later Lehman declared bankruptcy, right? It was terrible for all of us in our personal portfolios.. From Cloudera's selfish position it was actually quite a good thing.. What happened was we closed the funding round; the money hit the bank account, Wall Street imploded, the steel door slammed down on Page Mill Road.. I mean if you knew how to turn lead into gold you could not raise venture capital for that.. It was just impossible..

And that meant we just had the field to ourselves for like a year.. Nobody could enter the market.. That gave us a playground basically to go do this stuff we wanted to do engage with customers.. It was a very interesting time.. So, Mike, when you in the summer of 2008, what was it about Hadoop that kind of - and maybe tell all the folks here exactly - go into a little more detail in Hadoop and what was going on in the Valley at the time and how do you - why do you think there was actually an opportunity to turn this open source project into the company that it is today? So a very brief technical diversion I want to explain kind of how this software works.. 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 compute 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.. But maybe one thing that - just as a historical perspective, Hadoop ended up becoming an open source project at Yahoo and Doug Cutting, and what was that - how did that open source community evolve around Hadoop? So I presume you guys know the meaning, but the idea on open source software is a bunch of volunteers and companies around the world collaborate by sharing source code and writing source code together to make the software happen.. Google had invented this idea and written some research papers on it and there was an engineer in the Valley, a guy named Doug Cutting, who read the first Google paper in 2004 and he happened to be working just in the way one does on a project of his own to index the entire internet.. And he recognized, he was not going to be able to do that with conventional systems.. He read this Google paper and he got excited.. He thought I can use these techniques in my project..

So Doug and one other guy, a guy named Mike Cafarella, who is the faculty on the East Coast now for - I'm going to blank, and I don't want to get it wrong.. Doug and Mike created the Hadoop project to turn Google magic into basically open source.. The rest of the consumer internet companies recognized that Google had an unfair advantage in search.. They were able to do

stuff that no one else could do.. And so Yahoo in particular decided that it had an opportunity to disrupt Google and to end Google's hegemony in search.. You can argue about whether or not Yahoo was successful in doing that, but Google embraced this open source project, staffed it up.. Facebook piled on and hired a bunch of engineers that began contributing to the project and this Hadoop open source software emerged from the consumer internet as a collaborative effort driven really by the competitive dynamic with Google against the rest of the industry.. The decision to make that software open source that was Doug Cutting's ethical and strategic commitment.. He believes in open source by the way, so do I.. I mean, a bunch of us at Cloudera have pretty deep open source roots..

This platform would never have gotten this good.. It would never have advanced this quickly if it had been owned by one company, if we hadn't been able to bring together the genius of the entire planet and then by the way if it hadn't been so easy to get and so easy to adapt to specific use.. So the open source decision was huge in establishing and advancing Hadoop beyond anything else in the market.. And did that drive the decision to bring Doug into Cloudera within months of the founding of the company? Yes.. Well, honestly - so Jeff and Christophe and Amr and I, one each from Facebook, Oracle, Yahoo and Google came together to form the company in the very beginning.. We were from the very beginning deeply interested in having Doug join the business, for all the obvious reasons.. I will say Doug is a very good friend and I like and respect him and I think he feels the same way about me and I will go ahead and speak for him now.. I will tell you that probably in late 2008, he's probably thinking, I don't know about these Cloudera guys, there might be a lot of companies that are commercializing Hadoop.. I'm not sure who these guys are, be a little bit careful, right.. I might have to go be Linus Torvalds and be like Switzerland in the Hadoop ecosystem..

So we didn't manage to close him in the very beginning.. I think he was really encouraged by Cloudera's demonstrated commitment to open source and our investment in creating software that we gave away as part of the community.. He got to know us generally and he and I in particular began to build a relationship and by August of 2009 we convinced him that he belonged in the business and at that point he joined.. I'll say that Doug lives in Saint Helena, California.. He has got a beautiful home and he and his family live there.. He grew up in that town, and he is raising his kids in that town and he has got a little 10 by 10 shed out the back of his house, where he makes open source magic happen.. And he mostly just lives and works up there.. So once a month beginning in January, I would drive up and I would have lunch or dinner with Doug to try to recruit him to Cloudera.. And you know I was just like not making progress.. I finally decided I clearly need to up level this conversation..

So the next time I went up, beginning like in about June, I started bringing my wife Teresa, and inviting Doug's wife Ann to join us.. And I just started recruiting the family, because I figured I'm not the perfect guy, right.. Yes and I will say that Doug like me, I'm sorry; here is why I want to say this.. Ann Cautrell Cutting would be married to the inventor of fundamental technology that changed the world and Doug would be pumping gas if things had worked out differently.. The two of them hooked up and Ann made the right decision for the family and Doug joined Cloudera.. That's a great story.. So, just you mentioned a word big data earlier, I remember when Cloudera was founded in 2008, big data was not a word.. Yes, you guys didn't know what that was.. Yeah, now it's on every billboard, it's hip to say big data.. Cloudera gets a lot of credit for kind of causing this wave of big data..

What does big data mean to you actually if you had to boil it down? 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.. So you've talked about how the Internet data centers have used Hadoop, web logs and clickstream analysis.. Can you give some examples of how Hadoop, now that Cloudera has got hundreds of customers, is the leader in Hadoop and big data.. Tell us how it's being used outside of the internet data centers and what are some examples of people working with it? Yes, I will give you guys a few examples that I particularly like..

Let me say that, I told you before I'm a database guy.. And I'd been working in the relational database industry for about

two and a half decades before we started Cloudera.. And you know over that time, I made a very good living.. And we had some really successful companies, but you guys, it is kind of boring, really.. I mean tighten the screws, speed it up 3% a year, raise the price 5% a year, worked out okay, but nobody ever felt like, ah man relational databases, this is what I was born to do.. Larry Ellison thinks that still but.... No, he was born to do many things.. Relational databases are merely a means to some money, which is a means to an end for Larry.. Is this going out live? Larry I love you brother.. So we make a very good living at Cloudera, helping companies run workloads..

And I will describe some of those workloads in a minute.. We sell to banks, we sell to insurance companies, we sell to hospitals, right.. But it is my deep conviction that as a society, we're going to attack really important problems in the next decade.. We've got to figure out how to produce and distribute clean water to people in a warming world, and we're going to figure out ways to do that by looking at data in new ways.. 7 billion people on the planet today, forecast to be 9 billion people in 2050, that's 2 billion new bodies with no available planet to find new farmland on.. We have to figure out a way to produce food more efficiently.. And we're going to do that with data.. Meaningful cancers will become manageable chronic diseases in our lifetime, will no longer be a death sentence, but will be diseases that, like AIDS in rich countries now, you're going to actually manage and live with for a long period of time.. And we're going to do that by understanding the disease by using data to understand its progress better than ever before.. Data really is going to matter to us as a society in ways that it hasn't..

IT is going to be a social good.. I believe this deeply and we actually see this happening in our installed base.. So while nobody woke up feeling that way about relational systems.. Look, I know we're not going to cure cancer at Cloudera, we're not going to feed 2 billion people, but we're going to make software that lets the planet attack those problems in new ways.. I'll give you some concrete examples of use cases we see, some fairly pedestrian, some a little more inspirational.. If you are a credit card provider, you care about fraud.. There are bad guys all the time stealing credit cards, using them on the web, trying to get goods and not get caught, that's always been a problem and those firms have been looking for ways to detect and prevent fraud for a long time.. They've been using techniques by the way like machine learning, at small scale, looking at the last week's or month's worth of transactional activity.. It turns out if you can feed those models a decade's worth of data, you can see patterns, you can learn behaviors that were invisible in small amounts of data.. And it is glib example, but flip a coin three times, you know nothing about that coin..

Flip a coin a thousand times; you get an idea whether or not that's a fair coin.. It turns out that lots of data yields value, disproportionally, right.. Smarter algorithms are good, but more data to train your existing algorithms even better.. One of our customers a global credit card - basically a global card processing company, had been doing fraud detection at a small scale for a long time.. They brought us in just because they wanted to store way more data, way more cheaply than ever before, LAN the last 10 years' worth of transaction data and that was for cost reasons, that's all.. Save a few bucks.. Once they did that, they had all that data on spinning disks, some of their analysts said you know what let's turn our fraud models loose on this data, tweak them a little bit, but let's feed them 10 years' worth of the data.. They discovered the single largest instance of fraud in the company's history as a side effect of saving money on storage.. So just aggregating data has allowed them to do an old thing in a very powerful new way.. One other example I will hit and then I will let you either direct with more or go on to another question..

So one of our customers is a company named Explorys Medical and they're based in Cincinnati, they're - I'm sorry they're based in Cleveland, they are spinout of the Cleveland Clinic.. Much in the news these days of course are Obamacare and healthcare.gov and the challenges we have around that.. And whatever your politics are, I think you will agree that better patient outcomes more reliably at lower costs that will be a good thing.. So you get sick, you go see your doc, you're filling a form when you get in, maybe whacks on your knees, takes your blood pressure, talks to you about your symptoms, writes up a script, sends you home.. You fill that script, you take the pills that you're prescribed for a week or two, you don't feel much better, right.. You go back, see your doctor again, maybe some more invasive testing, blood draw or other body samples happen, maybe some imaging, maybe calling a specialist, look at another doctor to give you some more feedback.. This time maybe you get some physical therapy, you get a different prescription, you follow that course of action, I'm rooting for you, so I'm going to say you get better in this one.. If you could string that whole story together, you'd have a pretty good picture of your illness, right.. The progress over multiple weeks, lots of different data types in there, lots of interactions, there is interstitial delays, the amount of time between your doctor visits, that's significant.. Now imagine you could do that for an entire hospital network's worth of patients..

You could use machine learning to realize that Dr.. Jones was getting disproportionately good outcome from women between 35 years and 50 years old, by following a specific course of treatment when they present with similar symptoms and you could use Dr.. Jones' outstanding performance to make recommendations to Dr.. Smith and others to adapt their treatment.. That is exactly what Explorys is doing.. They're looking at the course of the care that aggregate - basically large aggregations of patients are getting and the outcomes, and thereby directing healthcare providers to get better outcomes faster.. Healthcare actually is going to be a big, big area for exploitation of data in this way.. So one of the questions I have always wanted to, switching gears, I always wanted to ask you and I've never had a chance is how .... I'm scared of this one.. ....

why did you choose the name Cloudera? It has nothing with to do with Hadoop, it has nothing to do with data and its - so

what was the story behind that? Yes.. Actually - so you guys should also know that my last company was named Sleepycat Software.. And what we learnt from Sleepycat was we don't let engineers name stuff.. Cloudera that name came about that summer when the four of us were scheming about creating a company.. At the time cloud computing was really a hot meme in the industry; we had no idea what our go-to-market or product strategy was going to be.. But it was reasonable to assume that some kind of cloud would be in there at some place.. And the era of the cloud well heck that's probably true, so let's go ahead and do that.. Significantly we could buy the domain cheap and we actually used a little subterfuge to - again, am I live? We did that in a clever way.. We really didn't have any idea what the company would do.. We chose the name because it sounded like a good name and because we thought cloud might make a difference..

Early on a lot of analysts and a lot of press misunderstood us.. They thought oh we are a cloud computing company.. Turns out we are a big data company.. In fact our strategy right now only kind of orthogonally intersects the cloud space.. Early on we got a lot of press ink because cloud was hot and we had cloud in our name and therefore probably worth writing about us.. It did have one bad effect and that is when you tie your brand to a meme in the industry that you don't control it kind of sucks.. So on balance I am satisfied with the decision to do it.. My point of view now is Cloudera is just the noise you make when you're talking about the most interesting enterprise software company on the planet.. And you think about Google, nobody really stops and thinks what a goofy stupid name Google is? Now you're like, oh yes Google.. That's a good answer..

So I think the - you mentioned earlier that you'd been in the relational database for decades before and now you talk about the power of Hadoop, but I've also heard you say that the relational database world serves a purpose, will thrive and live for decades to come and how do you see the world evolving in a database world having been in it for so long? So, let me first say we have been doing this for five years.. We have been able to watch customers adopt this technology and evolve the way that they use it.. Not just as their use evolved, but over the last five years we've been driving more capabilities into the platform, started out as this batch mode scale out storage magic from Mountain View.. But over time we pushed security, data governance, data lineage, the right kind of real time interactivity into the platform.. So the platform has gotten strictly more capable and users have gotten strictly more sophisticated in their use of it.. So while it used to be kind of off on the periphery, supporting new analytic workloads, but not really overlapping at all with your data warehouse or your document management system, it was a side system.. As it's got more secure and more real time, as customers have evolved their use of big data it's kind of moved in to the center of the datacenter.. We actually see it now as sort of an enterprise data hub, cheap, easy, store anything you like, full fidelity for as long as you like it secure, so you don't have to worry about it losing data, you can keep track of that data, you have got lineage and governance and it connects to your data warehouse, to your relational database transaction processes, your document management system.. That connectivity, not just to systems by the way, but also to users with the tools that they like makes it a really attractive place for data to go first.. And some of that data doesn't ever need to move out..

You're able to get at it in interactive new ways, but the products that I spent my prior career building are outstanding.. Online transaction processing systems grew up alongside that mission critical business workload for decades and they're unbelievably good at it.. High-end enterprise data warehouse, build a cube, spin the cube, fly through the cube, turns out that's pretty hard.. And those systems are really good at it.. What I think is going to happen is those high end domain specific services will remain as important as they've ever been.. But increasingly we will see data and other workloads move to this hub, because it's going to be a cheap and easy place to run that stuff.. I'm long-term bullish on relational systems, on data warehouse, I think that they will continue to improve and evolve and they are so deeply knit into the existing enterprise infrastructure that they will survive for a while.. I do think though that there is going to be a bit of a reordering and restructuring of where data lands first and how we choose to think about it, I honestly believe today and I wouldn't have made this forecast back in 2008.. I honestly believe that most of the world's data is going to live in this platform.. It's just too obviously a right architecture not to win in the long-term..

And just shifting gears a little away from technology into the business at Cloudera.. Cloudera is an open source company, but the open source business model has evolved significantly over the last decade and you've been part of that at Sleepycat and advisor to a lot of other open source companies, how would you describe choosing the open source model early on for Cloudera and what is the open source model for Cloudera or business model in general? How does open source play a role in that? Yes, I think it's a really important point.. And for those of you that are looking at careers in technology, what I will tell you is it's very different today than it was even in the late '80s, early '90s when I was first coming out.. Open source has happened in a really important way.. In the early '90s when we started Illustra out of Berkeley, there was no evidence that you could make money on software that people gave away, just flat out no evidence.. And the idea that that will be possible was ridiculous.. Over time a bunch of different strategies for monetizing open source have evolved, have been tried out, have been improved, some have been discarded, the world is now pretty sophisticated about open source.. What are some of those models, just as context? Let me, but I wanted to make one point.. Back in the early 19th century businesses actually had strategies about electricity, where were you going to get it or were you going to produce your own, now you know what you don't have an electricity strategy, it's just part of the landscape.. I believe open source software is the same way..

It was worth thinking about as a really discrete confusing complicated thing in the '90s.. These days most businesses touch open source in fundamental ways.. So if you want to be a business and you want to adjust open source software, heck if you want to write and give away open source software, how do you get paid? A bunch of different things have been tried.. One is take an open source project that is not very good and make it your own secret internal project and make it better, but don't

share any of your enhancements and really that's what we did at Illustra.. We took the Postgres code out of Berkeley and made it way better, but made it a proprietary product and didn't share it with anyone.. Sold it for cash money, just as if we were Oracle or Ingres or other proprietary database vendors.. And the problem with that is you surrender the contributions of a global community.. Your company has to all of a sudden be better than the rest of the planet in the aggregate.. Other models that have been tried include, well, you know the last version is freely available in open source, but the current version isn't, so version lagging.. You pay for the current version and the old ones are free..

You can do what MySQL and others like JBoss and Red Hat actually gets some credit for doing this.. I think Red Hat's strategy is more nuanced, but just be a services company.. The software is free, but if you want support, if you want consulting, you're going to come to me, you're going to pay me money.. The problem with that business model is that services companies don't drive real high margins and it doesn't leave that much money to invest in making the core product better.. So in general, those projects tend to languish.. Red Hat's strategy was actually a little more nuanced than that.. Red Hat offered a cloud-based management service, Red Hat Network, in order to operate those systems back in the day.. That was a proprietary software of theirs, it just happened to run on their data center machinery, it didn't get shipped out.. The operating systems they ran for their customers all of that was freely available, but part of their commercial value was in proprietary software.. Sleepycat; last example and then I will talk about what we do at Cloudera..

Sleepycat used a technique called dual licensing.. So the idea here was we have a good piece of software, but it's only good if you combine it with software that you have.. So if you're building a mail server or web server you'll want to put our software in it because then you get good data management services.. Well you can do that for free as long as all your stuff is also open source and free.. You got to give your IP away as well.. If you don't like to do that, well you can come and pay us money and we will sell you a different license.. We called it dual licensing, kind of viral open source for anybody that wanted to do that and if you didn't want to give your IP away you paid us money.. Another way to think about this is, it's little more jaundiced now that I have the experience of doing this, it's kind of like distributing the poison and selling the antidote.. The good business if you can get it, but your relationship with your customer begins based on a threat and that's not a really healthy place to start out.. So here is what I believe and here is what we believe at Cloudera..

Open source innovates faster, spreads faster, does great work faster than any single company can and embrace of the Hadoop ecosystem, giving away software as part of that ecosystem is really important for us.. But we have to have reasons for companies to buy our product uniquely and it can't just be that our people are the most awesome, because it turns out there are way too many awesome people in the world.. We're not going to be able to control that precious resource forever.. We can, we do build proprietary software of our own that lets our customers get more value out of their data, out of their infrastructure, operate it better, manage it better, secure it better, that's allowed to be proprietary from us and it creates differentiation from the pure play open source distributions that are available from some of our competitors.. It gives customers a reason to come to us; it gives us a recurring revenue stream that we can then invest back in the open source.. We call that really the Cloudera model.. It is I think the wave of the future.. I think it's a sustainable way to build an open source business and the best example I can give by the way is this is in some sense what IBM does with Linux.. It contributes substantially to the Linux ecosystem, but it builds and delivers database software and middleware and other stuff that runs on top.. That allows it to invest back in that infrastructure..

Anyway it's certainly been good for us: we're - you pick a metric number one in the market, but certainly revenues and customer growth and that's allowed us to invest more and faster and drive innovation of the platform in the way I described.. Sorry, I started doing a commercial there.. No, no.. Its - I think it's a very distinctive model.. I think one of the questions is, is that how you think it needs to happen for open source derivative software companies to be standalone independent, because historically they've all done the support services model, lot of distros they're acquired by you name it, big company and then the project kind of languishes? Yes.. Look what happened to MySQL, look what happened to Sleepycat? The fact that we didn't have a defensible IP strategy meant that at some point we couldn't continue to grow.. We would attract low-cost services competition, from outside the US, from cheaper geographies, just from people who recognize there was an opportunity to come; we would in some sense create our own competition.. I think that the only way to build a long-term independent company is to own proprietary IP.. I do think open source is the way that platform software is going to happen.. So if you're doing enterprise infrastructure these days and you think you're going to create a proprietary software company pure play, I think you're wrong..

CIOs want open source infrastructure, they insist on it and the ecosystem is capable of producing it.. If you want to build a company that doesn't get acquired by Oracle, but that grows up to be a company of that scale, really your only choice is to adopt a hybrid strategy like that and Red Hat is my favorite example, $1 billion in revenues plus now, continued growth driven off of a hybrid strategy that gave them unique value in operating systems and delivering open source to their customers.. That is our ambition for sure.. I think pure play services-only open source companies can and do, get to somewhere between 20 million and 50 million in revenues and there are many examples including our last company, my last company.. But for companies to grow past 100 and towards a billion, you must have a more nuanced IP strategy than just we're going to give all our bytes away.. And one of the good things that - because you've taken this model, Cloudera has created this really big and huge market.. The bad thing is when you create a big and huge market there becomes a lot of competitors into this pace.. How do you think about dealing with competitors, differentiating and how has that evolved over the last couple of years? So one thing I will say flat out and this is honestly true and from the heart.. We believed from the

beginning creating the company that we would be building a company worth billions of dollars, selling into a market worth tens of billions of dollars a year.. And there are zero examples of a single company in a space that gets to that scale..

The only way that you can make that happen is that if there is a rich ecosystem of companies building and delivering value in that space.. Relational databases would never have become big if it had just been IBM System R, right.. You need lots and lots of companies innovating and driving into the market.. So while the competitive pressure create new challenges for us and force us to up our game.. If we didn't have those challenges, you would have wasted your series A investment.. The entrance of players like EMC and Pivotal, of IBM into the market, much less the venture-backed little guys, that means there is an opportunity.. We have more than 700 companies in our partner program building apps on top of delivering integration services for shipping the hardware that our platform runs on, that ecosystem is focused on winning new business and driving adoption by customers, that creates huge opportunity for us.. That allows us to sell into a way bigger market than we could ever have made happened on our own.. We just got a lot more hands banging on the drum than we ever could have created on our own.. The emergence of real competition has forced us to up our game and you've been in the board room and very supportive as we've done it, but what five year old company is simultaneously in a knife fight with IBM, SAP and Teradata? I mean first world problems, guys, but problems, I mean, we are forced now to operate at a scale and with a professionalism that is unusual in the market..

It's driven by the scale of the opportunity.. We wouldn't have those competitors if they didn't see the dollars happening, but growing up that fast has forced us to be very, very disciplined and to be careful about bringing in new leadership and new capabilities steadily over the life of the business.. Right.. So I want to leave some time for questions.. I got a few more here for Mike, but if you guys have questions raise your hand and we will work them in.. So as you're thinking about your questions, Mike one of the - kind of switching gears a little bit to the entrepreneur, entrepreneurship and starting a company, what kept you up at night in the first year at Cloudera? And how is it different to what keeps you up at night today at Cloudera? So Ping makes an important point here.. If you're regularly sleeping through the night as a founder of a tech company, you're doing it wrong.. Seriously, there is always something that should be scaring the crap out of you, always, always, always.. In those very first days what we knew was we wanted to be a product company, we hadn't yet figured out what the product was going to be.. We had engaged with a handful of very valuable clients and we were struggling to try to abstract their requirements and figure out what our unique product strategy could be and in the experience of servicing those guys and delivering software that was valuable to them, we ran into real operational issues, they were running meaningful jobs on the platform, we had to service, we had to support them and we had to innovate on the product and figuring out how to balance that activity to have enough engineering capacity to make them successful, but then also to make our product successful..

That was a huge stressor in those early days.. That was the single biggest challenge, what will our product be? Honestly, I will tell you there is no question in our minds now of the scale of the opportunity.. We'll never stop being strategic in thinking about our product and where it needs to go.. But over the course of the next two years, frankly our challenge is going to be living up to the competitive set that we've attracted to the market.. I mean, I'm out there punching it out for deals with big enterprise vendors, who for two decades have been selling $70 million annual deals to C level executives of the Fortune 1000.. They visit each other at Christmas time in their homes, right, the salespeople and the execs and somehow we have to inject ourselves into those relationships as a bold little 500-person company based in Palo Alto and be credible.. And growing up that way requires discipline, requires excellence, requires new ways of behavior that we are learning, that we are adopting as we go.. And look I'm bullish.. I believe we are going to do it, but man I'm sweating that one, I'm sweating that one.. Any questions? Well, one thing is everything I've read says that credit card companies don't pursue fraud..

It's easier for them to just let somebody take a ding on their credit report than for them to go through the legal process of pursuing fraud.. The other thing is about cloud computing.. You're going to tell me that my data is safe with you and this is in an era when Sony loses 90 million customer accounts and the NSA has full cryptographic access to Google's data centers, I don't know, somehow I can't get behind this cloud data security.. So let me repeat and then I will address both those questions.. I will do them in order.. So the first question is - but why would credit card companies care about fraud, why do they chase that? I think they just let you take a ding on their - on your credit report.. Actually it turns out the law is written, so that in a fraudulent transaction you're responsible for no more than $50 of the charge and the bank eats the balance.. There is a huge incentive in credit cards to manage fraud.. Your debit card is another story.. Actually I believe that the incentives in the credit industry are very much aligned with fraud detection and we see that across all the major credit card companies..

They're all aggressively adopting and using new techniques to chase fraud.. And that's true broadly.. I mean when it's a cost to the business or when there is reputational risk; people want to get better at it.. So we absolutely see that happening and it's deep and real and it's not window dressing they really care, because it costs them real money.. The second question was - so, cloud data management and cloud storage, Cloudera, why would I put my data in the cloud when the NSA can steal it and when Sony can't keep track of it, and I will say couple of things.. So first of all, once again we just pick the name man, we don't actually store your data for you, we deliver a database system, that's used sometimes in data centers, sometimes in the public cloud.. But you do ask a nuanced question as large scale distributed cloud computing becomes available and as more and more data about all of us becomes available, how do we think about security, how do we think about managing it? A lot of companies freak out about the idea of putting their data in the public cloud because they think then it is at risk.. I'll tell you guys just flat out right now, right here, I bet you Amazon is better at security than 99% of the companies on the planet.. So if you're concerned about security, you should be much more concerned about your own internal IT team's ability to detect

and prevent break-ins, than you should be about the big guys' ability to do that.. They're much better at doing that given the scale at which they operate..

But the proliferation of data and its importance to the way we all work, I mean your medical data, your credit data, yes there are big questions there, there are important social, important ethical discussions for us to have about what's okay to capture and what is okay to do with that data.. I will tell you however that this genie is out of the bottle, okay, big data has happened.. It will not happen that we decide all of a sudden we're not going to do electronic medical records.. Ship has sailed, that's going to happen.. The next question is how do we secure them? How do we hold responsible parties involved in their dissemination? What is it okay to deduce from that data and what is it not okay to deduce from that data? If I'm an insurance company, should I be able to look at your DNA data if I may thereby be able to surmise that you've got a disease that might increase the risk of mortality.. As a society we need to have those discussions, but that data will be collected and is useful for important meaningful - I believe we are going to be better as a society with the ability to capture and process that data.. We need to be sensible about what we do with it and we really as a society, we haven't had that discussion yet.. What else? So in the past five years Cloudera has become like the de facto Hadoop distribution and then there's a bunch of other different ones that are coming up, Hortonworks, MapR.. What would have to happen, like who would have to come in to the market, for you guys be really shaking in your boots? So the question is Cloudera started first, Cloudera is the number one Hadoop company in the world, there are emerging competitors, MapR and Hortonworks were specific ones cited.. What would it take for Cloudera, what would have to happen to the market, who'd have to enter the market in order to really freak us out and really make us quake in our boots? Well there is a big collection of established enterprise IT players in the market right now and they have very deep pockets and the ability to plan on very long timeframes and I won't say we are terrified of those folks, but we are deeply respectful and attentive to the landscape in which we play right now..

And look another hint if you're going to go start a company, if you're not constantly kind of gaming out, okay if big company A acquired small company B, what would that mean to us? Or if small company B and small company C formed an alliance or if a disruptor entered, you need to be thinking about that all the time.. I think we have got a reasonably well articulated strategy that lets us not be reactive, be proactive, but wall off threats where we see them coming up.. So what would it take to have me quaking in my boots? If I could answer that question you should really find another guy to run the company.. If we knew the - if we knew of something that was that big an existential threat, we would address it proactively.. Hadoop came from nowhere to threaten a bunch of existing players in the market.. Could that happen to us? I pay a lot of attention to what's going on in innovation and data management right now, given where we are and we work very hard to embrace that innovation where it happens.. Is it possible that we could be disrupted in that way? Any given year the probability of that's pretty small; over 30 years, it's a virtual certainty.. I don't spend a lot of time worrying about an out of left field, never could've predicted it, two guys in a garage made a new thing happen.. We pay a lot more attention to what are the big vendors doing, what are their strategic interests, what are the emerging players doing, what alliances there might disrupt us and how should we position ourselves to be best able to deal with whatever happens in the market.. Through the various companies that you started and working with a lot of different types of people when creating Cloudera and in the other companies you started, what do you think is the most important thing that you looked for in the people you surrounded yourself with that allowed you to really push one another forward and to work successfully together? So the question is over the course of my career in a lot of different companies, what has been the most important trait in the people that I have surrounded myself with in order to build success? Let me say just for the record, Britton Lee that one just kind of petered out, wasn't a crashing failure, wasn't great..

Illustra was a tremendous success until the acquiring company restated earnings three times in two quarters that sucked a little bit.. My next one year mulligan I won't even name, but boy we flew that one into the ground at very high speed.. Sleepycat was tremendous outcome and then Cloudera book's not over yet, but so far looking pretty good.. I'll talk both about Sleepycat, but even more about Cloudera.. The quality of the people in the organization that's really all that matters.. In my view everything that we can claim as a success ties to the fact that we have recruited the world's best team and say that right out loud in front of everybody, I think we have been fantastically successful in recruiting a world-class organization including just lights out fantastic developers.. We have a very deep and careful process for sourcing and filling, sourcing candidates and filling positions and we interview not just for skills, we interview for capacity to learn, because the market is changing and we've got to get better as time goes by.. I'm also really - when we are hiring people in the company, I'm looking for people with multiple diverse interests, not just I am a fantastic C++ coder, but I'm a fantastic C++ coder who has summitted five of the nine highest peaks on the planet.. Someone who does that has a suppleness of mind that you want.. They chase tough challenges and you can rely on them to be interesting..

One reason I do that is I spent a lot of time on the road and I travel with a bunch of business colleagues.. After five days of 10 hours of customer meetings, on Friday evening when you're sitting down over a beer after that week, you want to talk about something and by God it's not work.. So this - the summitted five mountains guy, you can talk to that person.. Anyway, we look for diversity and intelligence and flexibility in the people that we hire and I will say, I think we have been pretty successful.. It's what makes Cloudera such an exciting place to work and if you're good at hiring that way early, it's easy to keep it up.. If you don't do that early, it's impossible to change it.. The reason is really great people gravitationally attract really great people.. Who wants to go join a B team and be like the A person? Yes, I got better things to do, but you walk in, you think you are the smartest kid in the class; actually you guys had this experience coming to Stanford.. Wherever you went to high school you were the geniuses right.. You show up here, you're not even the smartest person in your row..

That's a good place to be.. That's a good place to be.. If you consider five companies like Cloudera, Oracle, SAP, IBM and Salesforce, where do you see the landscape in five years and then if it's different in 10 years? So the question is if I look at Cloudera, SAP, Oracle, IBM and Salesforce, where do I see - what do I see happening? So let me say a number of those companies are very good partners of ours and I'm not going to prognosticate on that.. I've certainly got a view, but I'm not going to prognosticate on it and I think the relationships we have in the market, and the opportunity for everybody is big.. We are certainly focused on growing our share and growing the market.. I will observe however that in that population, Salesforce has of late been a breakout dramatic success and the reason is not that it has been wildly profitable and in fact it's never been profitable.. That's a business that's successfully lost money like forever.. But it is driving unbelievable growth on a recurring revenue stream that is to public market investors and think about long-term value of the company, a very large installed base that keeps coming back, that is a tremendous place to be.. When we think about this opportunity, what I'd love to do is grow as fast as faster than the market, that sort of growth as Salesforce has proven other companies have done likewise Splunk, Workday have proven that you can live that way, you can fund that growth, you can deliver huge rewards to your shareholders in that way.. Of course all these businesses one day need to pay their bills the old fashioned way with profits..

So no question you need to get there, but as markets are exploding, thinking incrementally is a waste of time.. We're down to just about two, three minutes left and we're have told that we have to stop sharp on the hour.. So maybe just one more.. So if you think of Hadoop as like one mountain that you can climb and there are these other open source projects, what are the ones that you see that are up and coming that look appealing or, and the other thing is tell us about getting your first one or two customers, what experience was that? So what other open source projects do I think are.... Could be open source in general, but I was thinking databases.. Yes, well we pay a lot of attention to stuff that bears on Hadoop.. Right now a couple of projects that we are deeply interested in: OpenStack for elasticity and sort of cloud deployment of this infrastructure.. We think private cloud's going to be huge and OpenStack has got a ton of momentum.. We just announced a relationship with a Berkeley spinout called Databricks and they're commercializing a project that was born at UC Berkeley called Spark for high-performance in-memory analytics of data.. You can imagine data analysis is really sort of in our wheelhouse..

So those are a few that we are paying attention to.. There is a ton going on in the open source ecosystem.. One of the best things about my job is I get to walk around, look at all of it.. But those are two in particular that we are paying attention to.. And then you asked as the second part of your last question, I will note, some quick stories about winning early customers.. Early on our investors Accel in particular helped with great introductions.. We met through them a big bank who were super interested in this new platform for data analysis.. And I'm going to tell this anecdote without identifying the bank for reasons that will become obvious at the end.. So they really loved what we were doing, they really wanted to use this new capability, but hey security, data privacy, they have a team called Information Risk Management that looks at vendors and their products and decides whether or not it's okay to use them.. And they have got a book you got to fill in, in order to qualify..

It's key to us to win in this space, so we want to take that deal down.. And we met with Information Risk Management at this bank like six times.. We had the same meeting six times in a row.. It was just like we are not making progress after six times I'm like, guys timeout, what's happening here? Why are we not - the risk management guy reels back in his chair, well, Mike, you know we put the no in innovation.. I go yeah okay, well, I guess we know where we're then.. Living in big company is tough and you've got to do it.. We are at the end.. Thank you, Mike.. Really appreciate it...