Divesh Srivastava Speaks Out

on the importance of looking at real data, abstracting problems and more

by Marianne Winslett and Vanessa Braganholo



Divesh Srivastava

http://www.research.att.com/~divesh/

Welcome to this installment of ACM SIGMOD Record's series of interviews with distinguished members of the database community. I'm Marianne Winslett, and today we are in Shanghai, site of the 2009 International Conference on Data Engineering. I have here with me Divesh Srivastava, who is the Executive Director for Database Research at AT&T Labs Research. Divesh's research interests include logic programming, OLAP, data streams, XML, data quality, and data anonymization. Divesh claims that he personally has no honors, but is the boss of all those AT&T researchers who've won many prizes, including best paper prizes at ICDE 2009 and VLDB 2008. His PhD is from the University of Wisconsin Madison. So, Divesh, welcome!

*Divesh, AT&T is a phone company – in fact, it used to be **the** phone company in the US. What do databases have to do with the phone company?*

I can say two things. First, AT&T used to be a phone company, but it is now much more than a phone company. Think of it now as a communications company, or a networking company. In the phone network, there is just one kind of network, and the operation of these complex

networks require a lot of data, right? In the operations, you can generate a lot of data, and to understand how the network is operating, you need to do a lot of data analysis. You need data analysis also, even for understanding how to design and how to evolve the network, whether it is the phone network, the IP network, the wireless network, right? And all of these things require a lot of data. So because of that, over the years, AT&T has had a history of collecting and analyzing various kinds of, I guess I should say packet-header data, for there are many kinds of networks, and because of that, there is a history of both collecting and analyzing large amounts of data. Trying to model what's in the network requires extremely complex databases. So in that sense, databases are almost critical to AT&T.

Then, maybe I should ask you what you can tell our academic audience about real databases that they don't already know.

I guess real databases tend to be very complex, as well as very large. *(How complex? How large?)* In AT&T, we don't necessarily have a single database for everything. They just evolve over time, so there is a lot of legacy, but we have even single databases that may easily have upwards of a 1000 tables, 30-40,000 columns in the database, and it is not your employee department database, it is not even a database where you can put the schema on a piece of paper and stretch it out over the floor of this room that we are sitting in, and even understand it at that level. So it is very complex. At the same time, we also have other databases that are extremely voluminous. So we have databases where we have close to a trillion records. That is huge. We have databases where we are collecting data in a streaming fashion, and it is so huge, that we can't even afford to store all of it. So that is some of what happens in AT&T as an example of a real database.

What about data anonymization – how does that tie back to AT&T's interests?

It is a recent interest. I cannot say that I have a complete understanding of how all it may be used for AT&T. But one of the interesting things, and maybe this is of interest to you, Marianne, given your interests as well, is often for reasons of regulation or law, we are all still… say, okay, we have been collecting this data for a particular purpose, for the network operations, but maybe it contains a lot of sensitive information. So you can only retain it for a certain amount of time. This certain amount of time might be 6 months, a year, 2 years, depending on the nature of the data. But given that we are in research, we aren't interested in necessarily violating anybody's privacy, right? But we do want to understand trends, and trends are often over much longer time periods than the 6 months that is needed for the operational use for that data. So I would like to be able to say, can I take, let's say, 10 years' worth of data, and anonymize it in a way such that it meets the legal requirements for the auditors, that we are not trying to preserve data that might violate somebody's privacy? But there are enough trends and statistics buried in the data that I can use it for analysis. So I would like to be able to anonymize data over a very long time period, keep it so that I can use for interesting analysis, that I may not even have thought about today. So that is one of the reasons why I think AT&T should be interested in data anonymization. But as I

said, it is a recent research interest of mine, and how the techniques that we are using today actually work to solve this problem? I still don't have a good answer yet. Is that fair?

That's fair! That's fair! You were working on streams before almost anyone else. Looking back at what we've accomplished in that area, where do you see that work having the most impact?

Certainly that is true among the database community. But there were people working on streams from the theory and algorithms community way before. Munro and Patterson's is from (I think) 1980[1]. And that is one of the first papers that talked about doing certain computations in a streaming fashion. And at that point, nobody had thought about streaming as a viable technique. But I think this goes back to one of the questions that you asked me earlier: AT&T has huge amounts of data, and can collect huge amounts of data, and because they realized that we cannot necessarily store all this data while still people are interested in analyzing this data all the time, that's why we got into streams in the first place. And so, where we got into streams, I think it was in 2000, 2001, around that time, is where we are in data anonymization today, right? We see a need, we begin to sort of understand how it might be useful, but over the last, close to a decade, we have done very interesting research in terms of core foundations for streams, we have also built data streaming management systems, probably among the most efficient there are. Because we have to be able to analyze data at the rate that data is moving through the network.

So has that technology made it into production systems at AT&T?

Yes, so AT&T's data stream management systems are tools that we have developed and are something that are being used on an operational basis in various parts of AT&T's network. It is used in some parts of the core network, and because of that we have to deal with networking speeds, that are like OC192, OC768, which are sort of at the leading edge, right? Try processing, something like a million, ten million, 50 million records a second…

Not with my hardware…

So in this case we are using, not necessarily massive souped-up hardware, but this is sort of well designed hardware, but also very well designed software. We are talking about this from a database point of view. Interestingly, this research that we are talking about (data stream management systems), really came about because there was a need in networking. And the database people had the technology, the ideas. So it is really a collaborative effort, between the database and the networking folks. And actually, that is a common theme at AT&T, right, we like to collaborate. We can't say that database people, sitting in their rooms by themselves can solve all the data management problems. Collaborating with the networking folks was critical. They couldn't have done it by themselves, we certainly couldn't have done it by ourselves, but happily, getting the two groups together, we did something.

[1] J. I. Munro and M. S. Paterson. **Selection and Sorting with Limited Storage**. Theoretical Computer Science, v.12, pp. 315-323, 1980.

So now you are working on data quality. Is that going to be the next hot topic?

I hope so. I think data quality is one facet of something else I touched upon, which is, we are dealing with extremely complex databases, and some of these complex databases are not complex on day one. They are complex because they have evolved over a very long time period. They are complex because they are trying to capture aspects of reality, in this case networks, which are extremely complex physical entities, and over time, the people who used to know about some parts of the database because they had applications that use that data have gone away. Expertise disappears. So nobody has that (sadly) complete understanding of complex databases. So I think one of the big challenges is really dealing with an understanding complex evolving databases. And data quality to me is one facet of that, because, over time, people have data, and it gets added, and when people don't necessarily understand it, it is very hard to know when is the data correct, when is it capturing all the aspects of reality that you were hoping it captured. And so we have been looking at sort of different ways of dealing with data quality.

At this point, with data quality, it is like the blind men and the elephant story. You know the story of the elephant? There are a bunch of blind people looking at it from different facets, and thinking they have understood it completely… That is where we are, right? I think we have at least a half a different ways of thinking about data quality. So one of the simplest ways is sort of a sense that says, if I can understand the semantics of the data, if I can abstract it, not by sort of gathering requirements or something from humans, but just looking at the database itself, and statistically analyzing it to see what can we say about semantics, from a statistics perspective, from a distribution perspective, and things like that. Maybe then we can abstract that into some kind of semantics, integrity constraints of some kind. And then we can say, well, things that don't match the integrity constraints, maybe they are potential quality violations. But the way humans work, it is much easier for us to say, here is what I came up with as a semantics, here is what I think is a violation, you as a human who is an expert on this domain, can you tell me if this makes sense? And it is easier for them to answer yes/no, rather than just asking them, can you sit down and summarize what is the semantics of this database? It is not feasible. So that is one way of looking at data quality.

Another recent way we are trying to understand data quality is, people used to think of quality as something where dirt gets into the database, right? And people have all kinds of ways of trying to prevent anomalies. But we are trying to sort of take that one step beyond and say let's try to understand what kind of updates we could make to databases. Not all of the kinds of errors that creep into databases can be captured via integrity constraints. Even SQL statements, as complex as you want them. Are there better ways of trying to prevent dirty data from getting into database? Maybe I intended to make a certain update, I made a mistake in my thinking, I specified something different, it doesn't seem to violate any constraints, it gets into the database, and now it's in there. Are there ways to sort of deal with that? So those are sort of positions of two blind men, I don't think we are anywhere close to understanding all the ways in which one can think of quality in data semantics. But in some sense, that is one of the grand challenges. We

need to make sure that the data that people have in databases is something that is of good quality, something they can trust, something they can rely on, because otherwise, people will stop using the data, and then it will deteriorate even further. So that is why I think it is important.

I have been told that you "think like a theory person", meaning that you are good at abstracting problems.

I hope so!

Don't people tend to define a problem to be what they can solve?

Yes, but hopefully, as researchers, we constantly learn, and constantly learn of solving things in new ways. It may be true that how I abstract things today is a function of what things I think I can solve today, or have some idea of how to solve today. But over the past 15 years, I've learned many things. I am sure that how I abstract things today and how I abstracted things 10 years back are not the same.

You think not?

I hope not, because otherwise I have wasted 10 years of my life!

Maybe, but you are abstracting from different problem areas. Now you are abstracting in data anonymization, and before you were abstracting in…

Sure! But, you know, being in a research community, constantly interacting with people… Again, I don't do this by myself. I am not in research because I like to sit in an ivory tower. I am in research because I like to collaborate with people. And I collaborate with new people, and they bring in new ideas, new techniques, right? So most recently, for instance, oh, not most recently, but as an example, I have been working with some people who have a lot of expertise in information theory. Ten years back, I did not have that expertise. Now I have a better understanding of it, so certainly, there are some problems that I abstract today that can make use of the information that I have, the knowledge that I have in information theory. Ten years back I didn't have that. There was no way I could have abstracted things in the same way.

That ties into something that several people mentioned to me, they said that although you worked with many different people and juggle many different projects, every time they meet with you, you give the impression that you have thought of nothing but their project since the last meeting. Several people also said that you pack 48 hours into every 24. So how do you do that?

Compression, I don't know! I enjoy doing research. To me, it is not something which is onerous. So, if you enjoy doing something, you think about it all the time. I don't turn off my research thinking. Was it Stefano Ceri who was saying he likes to think when he jogs? In the invited talk today? Well, I like to think while I am cooking. I like to think while I am walking. But I think, maybe the question came from people who actually do far more work on the projects than I do,

and I just provide maybe an insight or two, which they missed for some reason, and that may give them that impression.

So, since you have worked in many different areas, can you suggest a good methodology for choosing research topics?

So I can tell you how I choose a topic, and maybe that will work for others, and maybe it won't. But I have the good fortune to be in a research lab, right, where I have access to both very smart academically oriented researchers and colleagues, but I also have access to the industry, right? So I have access to people who generate the data, so I don't need to necessarily invent or make up applications. And having access to data, right, I think that is critical. There is no way I could have done a lot of the data streaming research or even the recent data quality research that I am doing without access to the data. This may sound a bit strange, given how we think of how databases are to be modeled, but I looked at some databases at some point in AT&T and we found database columns in a particular table that in the same column you may have a phone number, email addresses, what look like login ids, right, and you sort of say, that is strange, right? That is not the way people design databases. But it turns out that the column was being used as a user ID column, okay? For internal AT&T employees, and this was a database that came about over a long period of evolution merging data that was being used for different applications. And some applications had been using somebody's phone number as an identifier, which it is, right? Somebody else had been using somebody's email address as an identifier, and you merge these databases, you create an integrated database, and suddenly you have all these kinds of fields of information in the same column. And so we thought of, you know, this is a problem where you don't design databases this way, necessarily, right? But you come up with something, and now you suddenly have a lot of heterogeneity in the database. Can you even understand techniques for identifying heterogeneity? So we wrote a paper about it.

But without access to the kind of data that ended up there, having to imagine this, and if I were to simply say imagine that this happens, I am sure some reviewer would come back to me and say, you know, why would this happen? But know, instead of why would this happen, I know it is there. So having access to data in some sense is an unfair advantage I have.

Well, I think being in an industrial research lab, you should have some advantages that come from that, it seems fair to have some advantages.

I agree, I am not complaining!

You mentioned to me earlier that Bell Labs, of which AT&T Labs was once part, has many distinguished alumni.

Including yourself Marianne!

Thank you Divesh! So why are there so many alumni, but you've been there your whole career?

How do I answer that? I guess AT&T/Bell Labs have been around for a very long time. And certainly, even the database *alumni* have been around for a very long time. So people started, I would imagine even in the 1970s, right? So certain people like Rakesh Agrawal joined there in the early '80s, Jagadish was there in the mid-80s… and a lot of very good people… And they stayed for quite a bit of time. Jagadish was actually the head of the database group before I became the head. And he was there probably close to 14 years. I don't remember exactly how long Rakesh Agrawal was there before he moved, but it must have been at least I would say 7 years, or something around that time. I was even there when people like Carlo Zaniolo used to be at AT&T. But in some sense, I think of it as there is a lot of good people, right, and over time, maybe their interests change, maybe some people wanted to try out being a professor, some people wanted to have an impact on a database vendor, and so they moved over time. And I have been fortunate in having a lot of good colleagues that I continue to work with, so I am happy where I am.

We see a lot of students interested in database research coming out of IIT-Bombay, which is your own alma mater. How did that happen? How did that tradition get started?

I guess when I came out of IIT-Bombay (this was in 1987), and I went to grad school, at that point, I personally didn't have any sort of set view that I wanted to do databases, and I hadn't done any database research at that point. So I came to grad school in Madison, and Raghu joined as a faculty that same year, in '87, and then I took courses with him, at some point, and got interested in databases. But what happened at IIT-Bombay, much later was, there was a concerted attempt at recruiting people back from the US, right, and I think they made the good realization that rather than get one person, one person, in sort of a dozen different areas, which doesn't give people sort of a community, they decided to hire people in specific groups, and one of the first groups they chose, either actively or sort of fortuitously, happened to be databases. So they hired Sudarshan, who was my academic brother, right. He was also a student of Ragu's, who joined actually Bell Labs, after graduation, but then decided to go back to India. So even though Sudarshan was from Madras, which is a different city, he went back to IIT-Bombay.

Then they hired, I think, Sunita Sarawagi and Soumen Chakrabarti from Berkeley, right, both very strong, very good, database, some IR, some web, they went back. Then more recently, Krithi Ramamritham went back. And so it became like starting from a nucleus. It grew, and at some point I think they will probably have the strongest database research group in India. And so when you have such a good group of people, who are active in the research community as well, it gives an impetus to new students from there to continue that line of thinking. I think that is how it happened.

I heard that before you left Bombay you had never experienced temperatures below 75 degrees Fahrenheit. Did you realize what you were getting into when you left for Wisconsin?

Not in terms of the weather, definitely. This might be sort of partly incorrect memories, but I think the brochure that the University of Wisconsin sent, and you can tell me if the University of Illinois does this right, but all the pictures in the brochure, the glossy pictures, was of sun shine, people sitting on the lake, people walking on State Street. There was nothing much about sort of snow or 2 feet of snow and people sort of huddling in jackets. So of course I had no idea, right? But I sort of enjoyed the first winter I was there, it was fun. It was something new, and the cold still doesn't bother me. It wasn't expected, but it wasn't unpleasant.

I spent a month this fall at Infosys in Mysore, and you look more traditionally Indian than anybody at Infosys. How did that happen?

So tell me how do I look more traditionally Indian?

Well, you have more hair, you have more beard, and then you've got the different kind of shirt, and all the guys there were wearing pants and shoes like yours, but then they were wearing western style shirts and had western style hair. So they've all gone western.

Well, I don't think that is any different. I don't know if you have heard this, but when I came to the United States, I think my hair was probably this size (shows a very small size with his fingers). (*Ok, that's how they are now.)* So I don't think I was that different in India. I just happened to say, okay, I can grow my hair, let's try it out. And I liked what it looked like.

What about the Indian shirt?

I think it is very comfortable. It is very pretty, and periodically my parents, when they come, ask me what they can bring for me.

So they are supplying you with …

So they are supplying me with these things, and I like wearing them, so I wear them. Am I more traditional? I have no idea. I think there is a range of people. Or maybe I am just sort of still in 1987 when I came from India, and people in India have evolved.

Could be…

Could be.

So people describe you as "sphinx-like", with an "enigmatic smile". They say that it is hard to tell what you are really thinking, and that you don't reveal your strong opinions. How has this been helpful in your career?

Is this true, from your interview?

Well, I didn't know you were like that, but when they said it, I could see it. They know you better than I do. So when they said it, I could see that that was consistent with what I know. Because, for example, I didn't know what things you had strong opinions on.

I don't have too many strong opinions on too many things.

Well, I don't know, they said that you do, but you tend not to reveal them.

Maybe that is why I am a manager today.

Is it a useful characteristic for managers?

Sure, at least in some ways, right, I mean, there are different styles. People collaborate with people. I think I like my interactions with people to be pleasant, rather than adversarial.

In meetings, don't you ever have to convince people of something, say that streams are an important direction for AT&T or something like that?

No, because very often they realize that themselves. With my colleagues, for instance, people with whom I work, they don't work for me, they work with me. And if I try to tell them what to do, they would get up and walk out, which is exactly what they should do. If I am not able to convince them as opposed to just tell them what to do, I don't think that would work in a research environment. We are all very head strong. Nobody likes to be told what to do.

I can think of some people in the community who don't fit that mold though. But I am not going to name any names.

Can you tell us about the Rattlesnake Ranch?

So, where did you hear that from?

Oh, one of your colleagues, I actually don't remember which one. But if I did remember which one, I wouldn't tell you.

It is one of my favorite restaurants in New Jersey. So I like spicy food, very spicy food. The Rattlesnake Ranch buffet was introduced to me by a friend because it has a dish called the Chicken Habanero, where it is sort of extremely spicy, right, in a single chicken dish, maybe they have a couple of whole habanero peppers, sliced up and cooked with it. I like spices, so I like going there. They also have a kind of Margarita, which is a Blue Curaçao margarita, sort of a bluish color. So I like going there. And it is sort of an out of the way of beat place. They have a sign I think on the entrance, which I haven't seen anywhere else, which says, men are not permitted if they are wearing sleeveless shirts. It has a strange character to it. The food is very good. It is close to AT&T. I like it, and I often take a whole group or the whole department there for a department lunch or something.

Do they serve rattlesnake there?

Occasionally, they also have alligator, ostrich, bison, all kinds of stuff.

Ostrich is supposed to be the next big meat.

It is very tasty. I tried it a couple of times. I can't say I am a big meat person. But I like it, it is nice. Isn't kangaroo supposed to be the next big meat now?

I didn't hear that.

I think the Australian government is trying to promote that. Who knows if it is good or not…

As long as we don't start raising them here. I can imagine. My garden is already over run by rabbits and squirrels, imagine if kangaroos where popular also.

Your barriers around your house wouldn't work, right, they would just jump right over them.

Good point. The rabbits are deterred by very small fence, but a kangaroo certainly would not.

Do you have any words of advice for beginning or midcareer practitioners or researchers?

Well, I think, you know, whatever is fortunate enough to be involved in, right. Being in a place where you have access to real data, right, I think is very important. Twenty years back, which is when I started my research career, sort of, it was hard to get access to real data, a lot of real data. Today, at AT&T, I have access to AT&T specific data, but I think it is not that difficult any more to get access to lots of real data. People get access to data from the web, you can get access to lots data, lots of real data on the web, people can build crawlers. So I think one of the important things if you want to do database research, which is both research and practically relevant, is to look at real data, and look at lots of real data, because you can find things in real data that you won't be able to imagine.

Among all your past research, do you have a favorite piece of work?

I guess even that has changed over time. I am very fond of some of the recent data quality research I am doing. But that may simply be because I am doing it more recently. I certainly was very fond of the data stream research I was doing. I think even at this conference I have a data stream paper, which is a very cute paper. It's sort of is a paper on data streams dealing with decayed computations over time.

What does decayed mean?

It means for instance, if you want to be able to look at a stream of data, and some of the older data is still relevant, but less so than the recent data, so the importance of a data item sort of reduces over time as it gets older. For the longest time, people have realized this right, so they have come up with exponential decays and polynomial decays and things like that. It turns out

that in a data streaming system apart from exponential decay, people proved the hardness of all the other decays. It is very hard to do heavy hitter computation, which is needed for doing anomaly detections.

So you have this mismatch, well the mismatch said was, people realize the importance of having decayed computations, at the same time, the theoreticians, and maybe I am part of that community, sort of went ahead and said, doing it this way was very hard. So what we did was we sort of flipped it around and said, maybe you are thinking of decayed computations in one way, and there are alternative ways of thinking about it. So we came up with what we called forward decay as opposed to what people had been traditionally thinking about as backward decay. And then we showed suddenly with forward decay, it makes sense in a lot of contexts. Thinking about exponential decay in a forward or backward fashion is the same thing. Thinking about polynomial decays in a forward fashion is often more intuitive than thinking about it in a backward fashion. But you could suddenly do decayed computations with exactly the same efficiency you could do undecayed computations. Interestingly, even for exponential decay, you could do things like sampling, like reservoir sampling. And people had been trying to do it, and they came up with only partial solutions. But just flipping the way you think about it, thinking about forward decay as opposed to backward decay, suddenly there was a complete solution. It is a very cute problem, just a different way of thinking, which resulted in very interesting results.

If you magically had enough time at work to do one additional thing that you are not doing now, what would it be?

Probably learn more what other people are doing. So, in AT&T Labs, we are fortunate to have colleagues in a wide variety of disciplines, networking, visualizations, statistics, software… take your pick, right, speech. I am sure that if I knew better what the other groups were doing, I could find more interesting ways of interacting, of combining their application needs with what the database people know how to do, or learn new things of how to do. If I had more time, I would certainly like to do more of that.

If you could change one thing about yourself as a computer scientist, what would it be?

As a computer scientist, it is very hard say. Not too much. I luck into databases, thanks to Raghu. I think databases is one of those fields which has only become more important over time, everybody's data needs is only increasing. I don't know anybody who comes and says, you know, my database needs to be smaller than what they were a few years back. So I happen to be in an area of computer science where the research problems, the challenges, the needs, are only becoming more important over time. I am happy doing databases, and hopefully will for a while.

Great, thank you very much for talking with me today.

Thanks, Marianne.