Big Data and MPL:
An Interview with Josh Lee of Informatica

Josh Lee is Director, Global Insurance Marketing, for Informatica Corporation. Mr. Lee has been working on applications of big data in the insurance sector for many years, and shared with Inside Medical Liability some of his insights on how the medical professional liability community can apply big data in multiple areas of the enterprise.

Inside Medical Liability: Are property/casualty companies using big data today?

Lee: Yes, both life and P/C companies are working with big data. Without getting into too many of the technical details, I can say that, in terms of storage and archival, one of the really big benefits is that you can take any type of structure of data that has some science behind it, and it will help you index it and structure it internally, for subsequent searching and analytics.

What’s typically happened with insurance in the last ten years has been that companies have realized that the big vendors that sell database and storage appliances, like Teradata and IBM, are just extremely expensive. So one of the ways that we see big data being used is as a storage and archival site, or as a staging area within data warehouses.

The second area involves taking a lot of unstructured data to look at things like telematics, data coming from sources like cars which provide geospatial information and vehicle speed, acceleration, and braking. The car is pinging the raw data to the insurance company every few seconds. This is a large volume of data, and its semi-structured data. This can be handled as big data, for a fraction of the former cost. So, there is a massive cost savings for any company that uses big data.

The third area relates to what they call the “Internet of things.” “Things” here is all that unstructured data from devices, unmanned drones, cell towers, RFID or whatever.

IML: Who develops the models or algorithms for sorting and making sense of the data?
Lee: It’s going to have to be an application that is doing the analytics. There is a technology like SAS that does hard-core analytics. And then you’ve got the statistical programming language R, which does predictive analytics.

Within the insurance company, professionals like the actuaries, risk analysts, and data scientists take extracted data, process it, and do all sorts of “crunchy” things with it, like Monte Carlo simulations, algorithms, and formulas. These are usually done on an ad hoc basis, using customized tools. Sometimes, though, they are doing this on a regular basis—say, every month—taking an extract of data, crunching on it, and doing some sort of manipulation of it for pricing, say, or underwriting calculations or actuarial models.

What’s done is different every time, and different for every line of business. Sometimes it has more of a claims focus, for instance, or on product development.

IML: Are the professionals who are doing this having to develop whole new skill sets, or just adopting a wider perspective on the business?

Lee: Well, yes and no. First of all, the math and the science involved are just part of the education and training of professionals, in the continuing education they get in meetings and conferences. But, really, there actually is a whole new skill set.

When you think about R, it’s a programming scripting language and tool set from an open-source university project about twenty years ago. So if they’re going to use that kind of tool, they’ll have to learn a new language, just as they did structured query language, SQL, for databases.

So there are some new skills needed. One thing Informatica does, as a data integration company, is manipulate big data at the source. Big data platforms use all sorts of new programs, with weird names like Pig, and Scoop, and Hive. Basically, Informatica insulates people from having to do those themselves by leveraging simpler user friendly interfaces. We work with the hard stuff behind the scenes.

IML: In time, will some of these new programs fall into disuse, as we gain experience with big data?

Lee: I don’t think so. That sounds like a glib answer, but look at what’s happened even in the last 20 years. Relational databases are still here, as are mainframes. Large data appliances like data warehouses are too. Big data is here to stay. Will it be replaced by something else—better? Maybe. But by that time, people will have a bunch of big data clusters. It’s going to be about integrating all that data together into a meaningful business answer.

IML: Over time, will some of the data variables people are collecting start to seem weak, and will others have more predictive power?

Lee: It really depends. One of the things we’ve been seeing in insurance is the advancing field of predictive analytics. A lot of what insurance companies had been doing is a lot of rearview mirror reporting. The predictive models now are getting a little bit better—in predicting new areas to find new business, or customer bases, or predicting new risks, or develop new products.

So I do see that those models are getting a little bit better. There are a lot of sources that are coming in that insurance companies are taking advantage of, like social data, like from Facebook and Twitter.

That may seem silly, but what that’s helping them do is understand their policyholder base a little better. Understand their sentiments and their desires—their thought process. Maybe that’s helping them market
better, maybe that’s helping them develop new products, or look at socio-demographic trends, economic trends. So yes, things are getting better.

IML: So it’s not solely a focus on quantitative data—it’s also a focus on softer information?

Lee: Yes, it’s all of it: hard data, and soft data, too. Soft data right now is typically being used by marketing departments. Marketing departments within insurance companies have some socio-economic data. So they’ll know an age bracket like 18-25, 25-40, male/female, income brackets. They know that stuff, and they’ve known it for a long time.

Now, they’re starting to learn a bit more, so they can look at the picture at a macro level. They can ask, are most of my policyholders conservative-leaning or liberal? What’s their ethnicity? What are their interests? That’s important from a product marketing and product development standpoint, where you look at the key demographics.

So all of it is pretty important. It will become more important when you look at how maybe, over time, they will be able to do an analysis like exploring the relationship between workers’ comp and broader social trends, or perhaps on individual person’s social pages that may be available for fraud in, again, workers’ comp or medical claims.

IML: Does your profession think of all of this as—I can’t think of a good phrase—“ethically robust”?

Lee: I know. I’ve seen some interesting presentations over the last couple of months by data scientists who were in the charitable space, the helping people space, and they were talking along the lines of Google’s old motto, "Do no evil."

But data is what it is, and it can be used for good or for evil. I personally believe that you should use data for good, maintaining some of the core principles and liberties we have as a society. That’s kind of a metaphysical, personal, moral type of a decision that everybody has to make.

IML: For every type of insurance company, is it the same roster of data that they’re compiling—or does it differ, depending on their market?

Lee: It differs, depending on the market. We talk to companies of all sizes, of all types. Life insurers differ significantly from P/C. Life insurers, because they’re diversified financial product mix, oftentimes seem more like an investment house or bank. P/C insurers, even among themselves, differ. You cannot talk about personal lines with an excess and specialty carrier. They’ll tend to tune you out. And there are very different kinds of data that are applicable to each.

But one of the benefits of big data, especially when it comes to commercial and specialty underwriting, is that a lot of what they have is unstructured data—it’s documents, it’s engineering reports. This should all be available somewhere to the organization, through the data channel.

IML: Does all of it have to be digitized?

Lee: This gets down to the technical aspect of this. The answer is, yes and no. The big data science does that for you. It takes what’s basically semi-structured data and learns from it.

Now, people are shoving new data into that big data cluster because it can take that semi-structured data. For instance, if I have a PDF document that has numbers and letters and so forth, I can put that right in there.
Then your Informatica integration tools can take elements of that document out, and migrate it into relational databases or data warehouses for additional analytics.

**IML:** How do you see big data being used in the P/C industry in five or ten years?

**Lee:** I think it’s going to grow. With companies like Oracle or Teradata, to put up a 3- or 20-terabyte data warehouse will cost you millions of dollars. But you can probably do that with a big data cluster for under $100,000.

That's where I see the growth is going to be—as people realize, hey, we're spending too much money on mainframes or large data appliances. We can switch to these more agile data clusters, where I can just plug in boxes for a couple thousand dollars. Storage is cheap. Memory is cheap. The science will get better. Maybe something will come along and change that in ten years, but I think the growth is going to continue. And there will be growth in the analytics, too—all the data that people shove onto these big data clusters—will just grow and make more analytics available, make additional searching and tagging and cataloguing available, as well as just all the new data like telematics and things like that coming in.

**IML:** Is there some discrete point where data can officially be labeled "big data?"

**Lee:** People call it big data just because of the historic side of that term. There could be just a single text field in there. It really doesn't matter.

**IML:** Some of the PIAA member companies are fairly small. What would you tell them to induce them to use big data?

**Lee:** First of all, like I mentioned, the price point being lower does make it palatable for smaller companies. But then, you have to buy additional technology to sit on top of it, to be able to access it and/or learn new programming skills to work with it.

I think for smaller companies one of the things I would recommend is looking at providers in the big data space that have cloud-based platforms as a service-type offering.

What you’re doing is you’re not spending money on infrastructure. You’re not spending money on IT folks. You’re not even spending money on hardware and software. You’re basically just taking a subscription. Someone else manages that, for you. Depending on the provider that you use, there can be tools that help you get data in and out.

That’s what I would say to the smaller companies; that’s why cloud technologies are so significant, because basically it’s just taking, whether it’s software as a service, or platform as a service or infrastructure as a service, it’s just taking all that stuff off your plate, so you don’t have to pay for it.

**IML:** I was just wondering, as we were talking, what aspect of all this excites you the most?

**Lee:** What excites me the most is to see people do interesting things with their data. And it's been that way for me that for nearly 20 years. I've looked at a lot of the data myself. I really like to kind of get in there, and look at a lot of it and say, wow, look at these trends—something interesting is happening here, on this chart or graph. And I think insurance is interesting because of the broad range of physical assets that it insures. That makes the data particularly interesting.

*Josh Lee is Director, Global Insurance Marketing for Informatica Corporation; jolee@infomatica.com.*