Expert Interview Series: Part 1 — Dr. Ellen Friedman Discusses Increased Flexibility in Big Data Tools and Changing Business Cultures

JULY 12, 2016 BY CHRISTY WILSON

In this Syncsort Expert Interview, Syncsort's Paige Roberts speaks with scientist, writer, and author of numerous books on big data, Dr. Ellen Friedman. The two discuss how Hadoop fits in the industry, what other tools work well for big data streaming and batch processing, and about Friedman's latest book in the "Practical Machine Learning" series, called "Streaming Data Architecture: New Designs Using Apache Kafka and MapR Streams".

Ellen Friedman is a consultant and commentator, currently writing mainly about big data topics. She is a committer for the Apache Mahout project and a contributor to the Apache Drill project. With a PhD in Biochemistry, she has years of experience as a research scientist and has written about a variety of technical topics including molecular biology, nontraditional inheritance, and oceanography. Ellen is on Twitter at @Ellen_Friedman

Hadoop isn't just one thing. There is Hadoop itself, with all its multiple parts, plus the whole ecosystem of Hadoop infrastructure and architecture.

In your experience, what are people using Hadoop for?

I wrote a book a little over a year ago with Ted Dunning called "Real World Hadoop". For that book, I looked at how MapR customers are using Hadoop. We tried to talk about Hadoop so that it wasn't MapR specific. MapR is a different kind of platform, so some of the workflows we showed were a little simpler than in standard Hadoop. But you could do the same work with other systems, you would just have to have more pieces. The goal was to understand why people are doing this, and not just what drives them to it, like marketing. What are their needs? In some cases, companies are just now starting to recognize their own needs and how Hadoop fits.

Why is that, do you think?

Hadoop sounds like a specialized thing, but Hadoop isn't just one thing. There's Hadoop itself, which has multiple parts. A larger issue is the whole Hadoop ecosystem, which includes the whole collection of tools people use.

In general terms, Hadoop allows companies to scale up to levels of data they might not have even considered using before. And, by allow, I mean that it makes it feasible; it's practical, and it's affordable. So, Hadoop is opening doors, not just to do the same things people have been doing, but now at lower cost. People are beginning to ask questions that have never been asked, applications they can do that they could never have approached, because you actually do need a threshold of data to be able to do that.

The other purpose, that I think is less obvious to people until they really start working with the system, is that the Hadoop style approach, including the MapR platform, opens the door to a new kind of flexibility. Now with the new ability to handle streaming data, that agility changes the way people can analyze and process data.

Part of what I try to do is to help people recognize that they have a different option in terms of flexibility. That means they can begin on the human culture side of their organization to rethink how they approach their problems. Because otherwise, they aren't using the tool to its full advantage.

Can you give some examples of the kind of flexibility you mean?

One example is, as people begin to use a NoSQL database, like the MapR DB … it uses the Apache Hbase API, so it's like a columnar data store. It actually has a second version, which uses a JSON document-style interface that works sort of like Mongo, so it really goes in two directions. But I think the basic principle is the same. … Those NoSQL databases (Apache, HBase, Mongo, and the others), you can put in more data, you can put in raw data, you can process data, you can do these early steps in the ETL, even if you're going to use a traditional relational database later, you're doing that early-stage processing, starting with huge data and going down to smaller data…

Aggregating and filtering.

Yes, and you can do it cheaper. It's better to hand a relational database the good stuff, right? So you do your early processing on your Hadoop platform, in HBase or whatever. It gives you a different kind of flexibility.

Another flexibility example is that, unlike a traditional relational database, you don't have to pre-decide every question you want to ask. You don't have to know exactly how you're going to use the data. Not to say you start with no clue, I don't mean that at all. But let's say I know I want to use it for this particular purpose, I'm going toward a BI relational database for example.

But you're still storing that raw data in HBase or wherever. You didn't have to use it and throw it away because you couldn't afford to store it. Now you can ask, "What other questions, what other exploration do I want to do in that data?"

You don't have to have a big committee to decide it. Because the expense of going back and asking a different question in Hadoop is different than saying I want to do an entire new line of analytics through a relational BI database. That's a big commitment.

That type of flexibility means, people can defer decisions. They can save the data and then figure out what they want to do with it later. On the human side, they need to transform their thinking in order to take real advantage of that.

How do you feel about Spark?

As new tools come along like Apache Spark, people say, "Well, hasn't Spark replaced Hadoop?" I say, "Well, some of the processing in Spark certainly is replacing many of the situations where people are using the computational framework in Hadoop, which is MapReduce, but it doesn't mean it's replacing all of Hadoop. It means it's replacing just the piece that's running in Spark.

Tell me about Drill. MapR has put a lot of emphasis on Drill, I've noticed.

Apache Drill is a very fast, very efficient SQL query engine that works on Hadoop, or without Hadoop.

What makes Drill different?

Drill works on some of these new styles of data, Parquet, JSON, it works on even nested data. And you don't have to do a lot of pre-defining of schema. It discovers schema on the fly. The other thing that's unusual about it, as opposed to things like Hive or Impala that are SQL-like, is that it's standard, ANSI SQL. So it connects the traditional people and their traditional BI tools directly into this Hadoop-based, big data world in a much more seamless way.

But, one of the great things about Drill, back to this theme of flexibility, is that … people say, "How fast do its queries run? How do you compare it directly to one or the other choice?" Well, it's very fast. It's not always the fastest, it depends on the situation … But really, to understand if it's a good tool for people, the question is not, "How fast did the race run once the starting gun went off and you ran that query," that's important. But how long did it take you to get to the starting line? Was that three weeks of preparation of your data to run the query? Well, with Drill, that three weeks may become an hour or two hours.

So, if I got to the race in two hours plus ran it in 20 seconds, versus three weeks to get there plus 15 seconds to run. Which was faster?

Not only do you get to the end of that race faster, but now you realize you can take insight from that first query and turn around in the moment and say, "Oh, now I see. I also have another question." You loop back and you ask that second question.

So when people say, "How fast is it?" I say, "One question you should ask is, 'how long does it take me to get to form the second query?'." Because you can begin to build insight on insight at the rate people sit down and talk, like a conversation.

They look at a result, have a cup of coffee, chat with your colleague, go back in the same day and take it to the next step. That's just not possible in other systems where it takes hours and hours or days or weeks to prepare the next query.

When it takes that long, your thinking is different. The expense of doing the next set of questions is different. So you say, "Do I really want to ask that question? Is it worth it?"

As opposed to following that train of thought and going, "Well what about this? What if we do that?"

Exactly. So one thing that I try to help people see is with Apache Drill, it's not just that initial convenience, which is huge. Also, you want to structure your teams, how they do their work, how they think about things, how they move the work forward, what their goals are, differently to take full advantage of it.

So, to go back to the beginning, the purpose of Hadoop is to let you have access to new and unstructured data, and let you have access to traditional data that you've been using before, but at larger scale, much less expensively, and, on the other side, let you start thinking in new ways of "save data, go back to it later".

You don't always know what is going to be the important question about that data. You see that in research all the time. When people do a study, they save what they think is important at the time. You go back and say, "If only we had asked people in the case study this question." And you can't help that.

I think Hadoop lets people start to use data in a much more fundamental and exciting way. But it's not Hadoop versus traditional ways … it's how you connect those together that optimizes the system.

Tomorrow, in Part 2, Ellen talks about Streaming Data — what she finds the most exciting about technologies and strategies for streaming, including cool things happening in the streaming data processor space, streaming architecture, metadata management for streaming data and streaming messaging systems.

APACHE DRILLAPACHE HADOOPAPACHE KAFKAAPACHE SPARKELLEN

FRIEDMANETLHADOOPHBASEJSONKAFKAMAPR STREAMSPAIGE

ROBERTSSPARKSTREAMINGSTREAMING DATA

AUTHORED BY CHRISTY WILSON

Syncsort contributor Christy Wilson began writing for the technology sector in 2011, and has published hundreds of articles related to cloud computing, big data analysis, and related tech topics. Her passion is seeing the fruits of big data analysis realized in practical solutions that benefit businesses, consumers, and society as a whole.View all posts by Christy Wilson