experience what's inside™

THE POWER OF ASKING QUESTIONS

NAVIGATING THE ANALYTICS JOURNEY

Bob Rogers
Chief Data Scientist for Big Data Solutions, Intel
JUNE 2016

Interviewed by Daniel Magestro, IIA Research Director

INTERNATIONAL INSTITUTE FOR ANALYTICS™

DISCUSSION OVERVIEW

Intel's Bob Rogers, chief data scientist for big data solutions, sat down with Dan Magestro, research director at the International Institute for Analytics (IIA), to discuss the power of asking questions when assessing an organization's analytics maturity.

When you step into an organization, what is your approach for determining the sophistication level of an analytics program or an initiative?

The first thing is to ask them some questions around what they are trying to do with analytics now. What questions do they have? What questions can they ask? And where is the data coming from that they are using? The next question is about the shape of the data: Are they using very carefully controlled structured data? Are they incorporating any kind of broader data? Is that data coming from different locations within the organization? Or to flip it around, do they have data in their organization that they are not able to use because of either organizational constraints or technical constraints? So initially, the questions are around what they are asking of their data. Then, once we have gotten to that, the next set of questions is around what kinds of questions they would like to answer.

Is that approach primarily to uncover background information? Or does that approach itself start to elicit their maturity and allow you to assess their maturity?

It absolutely allows us to assess their maturity, because if they're not sophisticated about unstructured data, they won't know that they need to model the data inside their data center. In the old way of doing analytics, a business manager said, "I need a report on X; I need to know how many widgets we have sold or what the year-to-date units and revenue are for a particular group by SKU and geography." The analyst then would go and do some queries and return a result. That all was based on a system in which the data is actually very carefully structured into the shape of rows and columns. Then those rows and columns are connected together into some kind of a schema. The challenge that the analyst often has is, "I would really be able to answer this question better if I had access to that data over there." And that data over there is probably being generated in another database, or it's owned by a different business unit, but it's data they believe could strengthen their analytics if they could bring it together. The challenge is that they are not able to get that data.

For years, analysts have expressed that the inability to access data was preventing them from doing the analytics that they dream of doing. So along comes big data and the Hadoop* infrastructure, and then of course its progeny of Spark* and Apache Flink* and NoSQL data sources such as Cassandra*. All of a sudden you have the possibility to assemble data into a data hub or a data lake and bring data from all the parts of the organization into one place, and it looks like it's going to be nirvana. So the organization then embarks on this journey with this idea of building a data lake, and the insights are just going to bubble to the surface of the data lake. And what they discover is that even though they thought that somewhere in the data was the thing that they were missing, they've just got a huge mess on their hands, because when you're using data in relational stores, especially an enterprise data warehouse, you have a certain shape and structure that is imposed on you.

The ability to put data into a data center without requiring a very specific relational shape actually does not mean that you don't want to have a shape in the data. So going back to your question, if I ask them, "How are you modeling this data, this unstructured data in your data center or in your data hub?" they usually give me blank stares. They don't know what I'm talking about. And that's the point where you know that you need to go back and teach them a little bit about data modeling and understanding the relationships between the pieces of data, which is a requirement to do analytics.

"In addition to the right technical build-out, it's critical to have people who can tell the story in words."

EXPLORING THE SCOPE OF ANALYTICS

How do you characterize or label the different levels of sophistication that you can observe through asking questions?

I don't have an official taxonomy, but basically it starts with early business intelligence (BI) types of capabilities. So you've got your islands of analytics where different groups have a pocket of data, maybe a local database. It might even be running on a single machine, a single desktop. They're able to do calculations locally, and the kind of question that they can answer is something like, "What is the year-to-date revenue for my group as of last quarter?" Then the next level up is where they actually have some automation, so they're able to load that data more quickly. Then they can ask the question, "What is the year-to-date revenue for my group as of last week?" So they've shortened the time frame, but it's still just focused on their group.

Then there is the **early BI enterprise data warehouse level** where they have started to aggregate and assemble data into a single repository for multiple groups, and maybe even multiple business units. So the question that they can answer is, "What are the year-to-date units and revenue?" So now we've extended the number of pieces of data that we can aggregate "for my business unit by SKU and geography, as of last week?" That's essentially a SQL query on a relational enterprise data warehouse. The next level up would include a question such as, "What are the year-to-date units and revenue for my business unit by SKU and by geography as of the end of business yesterday?" At this point, we have brought it up to its current data, but it's very restricted to specific pieces of information that we're dropping into our data warehouse, and we don't have any agility around that. That's about where more than half of the organizations are, in my experience.

Then we get to **advanced analytics or advanced operational analytics**, and this is where we start to use more sophisticated techniques around data; we're also able to do historical analysis, which means that we can start to ask a question like, "What factors have historically influenced the units and revenue by SKU for my business?" And I can further ask, "Who are my customers, and what do they look like?" So I'm starting to get a view of not just what happened in each quarter, but actually what I think might be influencing the changes that I'm seeing and how that relates to who I am selling to. And then I actually branch down from there.

There is a subset for incorporating **Internet of Things (IoT)**, which is its own question. For example, a logistics company might ask, "Where are the packages?" This is the company's ability to know more or less in real time where the packages are in the system. So organizations that are at that advanced analytics stage have those timely analytics capabilities so that the data is up to date and the analytics are beyond pivot tables to probably some form of visualization. You've also gone to more factor analysis to understand what might be influencing things. This stage of analytics sophistication is all about cause and effect.

Then the next piece is **predictive analytics**, where you've not only got your basic business data, but now you've begun to be able to collect context data that tends to be unstructured. So now we're starting to get into the realm of analytics on timely data that includes unstructured data like what people are saying, what's happening on the Internet, or call center activity, which allows you to understand not only what the hot topics are, but also whether things are positive or negative.

At that stage—the predictive analytics stage—they are typically answering questions like, “What unit and revenue sales by SKU should I expect next year?” So it’s prediction. How accurate is my prediction? And that’s a very, very important point. Organizations that are at the level of predictive analytics can’t just be forecasting, which is something that has been done for years, even in early BI types of environments, they’re actually understanding what the limitations of the prediction are and what factors would influence it. And then they also can answer questions like, “What are my customers saying about my products and those of my competitors?” That’s really taking into account a lot of different kinds of information.

Most enterprises are definitely not at the predictive analytics level yet, but that is the question that they dream of. So in most of the conversations I’m having, they’re able to tell me some factors that have influenced their business, and they would really like to know how those factors are going to change what they should do in the future.

And then, finally, there are two more categories: prescriptive analytics and cognitive analytics. **Prescriptive analytics** is something that people who are in the analytics media are talking about a lot. The idea of prescriptive analytics is that the analytics actually drives decisions in an automated fashion and gives you information on what you should do. The question would be, “Given my forecast for units and revenue, how should I prioritize new product development?” That is, “I’m now knowing something not only about what I think is going to happen, but what I think is going to happen based on different products and gaps in my products, and then how can I most effectively reach my customers?” This is analytics that is filling in the blanks in my future planning. So not just what I think is going to happen with something that exists, but also actually telling me where there might be gaps in my current planning and helping me prioritize how to proceed with different options.

I think the true value is to ask, “How can we augment our human decision making with information from predictions and simulations?” And then the IoT version of that is, “How can I manage my fleet? What routes, what vehicle capacities, what fuel types, and what driving patterns would I influence so that I can optimize something such as lifetime of my vehicles or annual operating costs or whatever it is that I am interested in?” So again, that’s really a simulation game—for different types of choices, what would the outcome be, what would change in the outcome?

And then finally, **cognitive analytics**, which is when all of a sudden, it’s not the people asking the questions, but it’s the analytics system asking the questions … and then answering them! So an example of that in more realistic terms might be for the IoT, “What measurements or sensors might I want to add to my system to improve my prediction capability?” What it really means in practice is more fluent interaction with humans for analytical systems and the incorporation of less-rigid pieces of information. Even when you get to predictive analytics, if you’re taking value from unstructured data, you start to have to incorporate how confident you are about something. And if you think about how humans do things, we don’t explicitly think about the probability that something is accurate. We just make a decision, or we come up with a couple of choices.

> “I don’t think there is any organization that shouldn’t strive for predictive analytics.”

THE ROLE OF TECHNOLOGY + PEOPLE

Is your questions-based approach centered on assessing the people more than the process or the technology?

Yes. One of the biggest questions I get is, "How do I take the teams that I have and morph them into the team that can help me answer the next level of questions?" Typically that means you've got some analysts who have different levels of technical experience. Some of them might have statistics or mathematical backgrounds, while others may have computer science or more specific SQL training—literally, they've been trained to do structured queries in relational databases. And based on the questions that they are prepared to answer, you can map the skills to the questions and then identify gaps or needs for training. The question often boils down to, "Who do I need to add to my team, or what training do I need to give to my people, to get them to the next set of questions?" If I'm looking at forecasting, making a prediction, and I need someone who can understand how accurate the prediction is, at least right now, you're going to need someone who has more of a statistics or data science background, because the calculation of prediction accuracy is really a statistical probability kind of question.

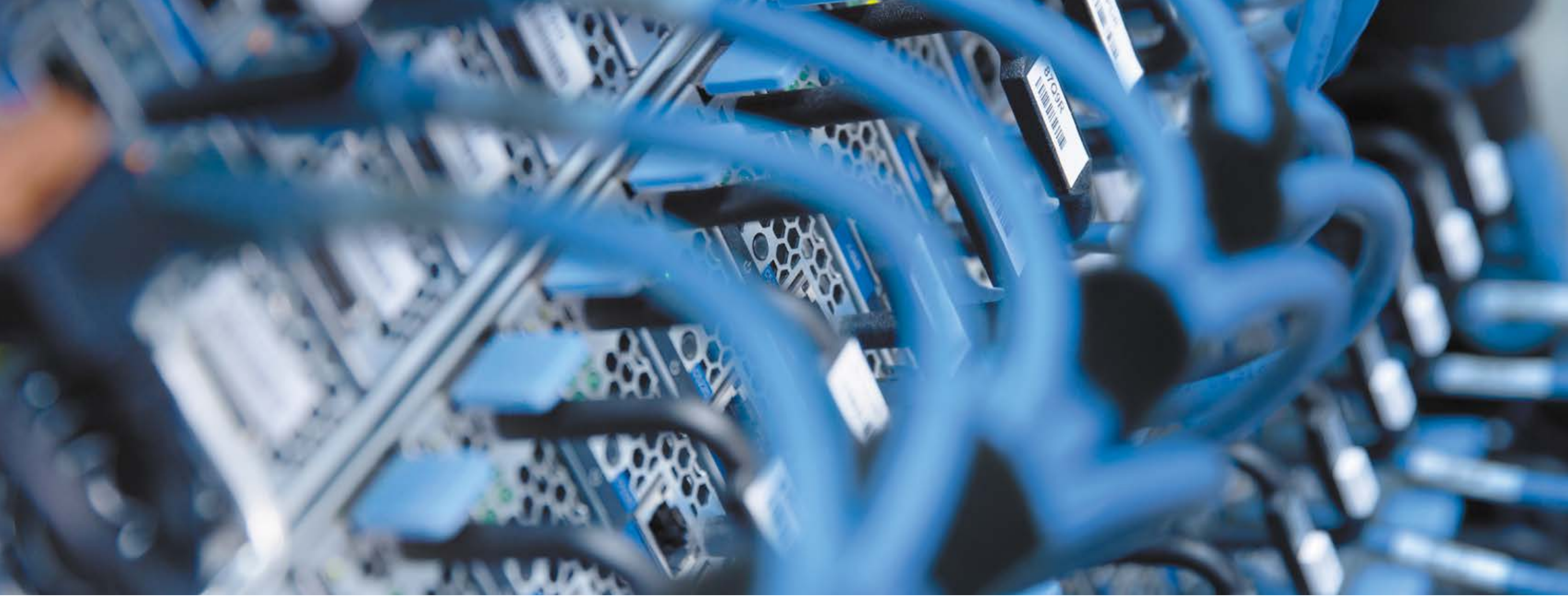
Often when a team is lacking experience in dealing with unstructured data, you really do want to hire somebody who can come onto the team and provide real, specific experience around handling unstructured data, because it is fundamentally different. You have to incorporate whether or not a piece of information that you have extracted is usable for the application that you need it for.

How does the usage of different technology solutions either fit into your assessment of sophistication or maybe hamper or trick your ability to assess sophistication?

That's a good question. Visualization is one of the places where the story gets tricky, because good use of visualization can actually help you get around some need for sophisticated analytics. One example is an old technology-driven business with a very dynamic leader who built up a data science practice within the organization. They built a cluster of 150 nodes doing a true big data infrastructure, and they aggregated and modeled their data. In fact, it was one of the examples that actually started out by aggregating their data without modeling it, and they had a huge mess. So they went back and restructured their system to clearly model the data as it came in, to understand what was there and annotate it. Then, rather than building out an underlying machine learning and analytics pipeline for scoring specific events that are happening with their customers, they built cubes of data for different product areas and trained analysts to use visualization tools to create interactive visualizations without the data.

In addition to the right technical build-out, it's critical to have people who can tell the story in words and connect that story in words to the analytics when necessary. I would say that is one of the biggest roles for an experienced data scientist on the team: to really teach everybody how to tell the human story as part of the analytics story.

> "The trick is to understand if an organization is connecting the data to the business objectives in a good way."



MOVING FORWARD

At IIA, we see that organizations often aspire to reach for their highest level of sophistication, such as machine learning and cognitive analytics. Is this the best approach?

I think the cognitive analytics piece is really a question about where turbocharged, humanlike capabilities could augment humans or usefully replace them. There aren't many environments where replacing humans with computers is a supervaluable activity. Most of the time, augmenting what people do with analytics is the sweet spot. And that lies more in the prescriptive, somewhere between prescriptive and predictive analytics. I don't think there is any organization that shouldn't strive for predictive analytics.

With prescriptive analytics, you're starting to talk about technologies that are very powerful when put in the right hands. So it's a question of maturity. Do you have the maturity in your organization to build out good prescriptive analytics or take advantage of prescriptive analytics in ways that are actually able to take into account the intrinsic uncertainty in what analytics do? So for instance, there are cases in which you can't really reliably forecast, and you need to be able to know when you're in that situation. You don't want to get into a situation where you're spending huge amounts of effort or making huge bets on predictions that are trying to predict that random 90 percent. And, of course, algorithms don't know about this. They just come up with a prediction. The most sophisticated algorithm comes up with a prediction, and if you're lucky, it might tell you some sort of error measure as well. But organizations that are striving for prescriptive analytics have to be aware of how applicable the analytics are for their business and, really, how valuable it is to use that.

On the other hand, having a measure of what is most likely to influence the future gives me a pretty good place to do some cogitation and at least a good place to start in my own decision making process. I think the answer is that absolutely everyone should be striving for predictive analytics. Almost everyone should be striving for prescriptive analytics on a reasonable time frame as they become more and more sophisticated with understanding the analytics and the limitations of them, and then pursue cognitive analytics as an interesting area for further research.

INTEL LOOKING AHEAD TO PRESCRIPTIVE ANALYTICS

Intel is playing an active role in the Trusted Analytics Platform (TAP) initiative, which is designed to accelerate the creation of cloud-native applications driven by big data analytics. TAP is an extensible open-source platform designed to allow data scientists and app developers to deploy solutions without having to worry about infrastructure procurement or platform setup, making analytics and the adoption of machine learning easier.

For some organizations, these kinds of questions may not be answerable today, but it might lay the path for answering them tomorrow.

Yes. And I think it's important to build for the next generation with an eye to the generation after that. So the questions are floating above you. As you transition from early BI to advanced operational level, you want to at least be thinking about what that predictive analytics target is that you might be going after in the future.

If you don't have that longer vision, it would be easy to say, "This is too much risk; we're not going to go to the next level," such as advanced operational analytics. If they don't have that broader predictive carrot on a stick on the other side of that obstacle, they're never going to get there. I would hope that if they see that predictive vision, then it will help them push through those fears about the existing infrastructure. And honestly, it's interesting that what they weren't really grasping is that their entire querying and reporting system could work on that big data platform. Without that vision, you can easily see how risk avoidance could prevent you from continuing on your journey into analytics nirvana.

In life, the valuable skill is not answering questions but asking them. And so it's been a pretty fundamental shift for me in the last 10 years to learn that if I'm asking good questions, then I bring a lot more value than if I'm offering just a few possibly correct answers. It's interesting, because I think this whole question context is actually very profound and really helps get through some of the confusion and the distortion.

Really, the trick is to understand if an organization is connecting the data to the business objectives in a good way. The first question is, "What is your strategy for identifying the questions that you're going to answer? And give me an example of one of those." And then given that question, "What data did you think was relevant, and what was your strategy for bringing that into the analysis?"

The nice thing about that is that it's completely technology-independent, right? It's about, "What is your understanding of what you can do with analytics and with how the different pieces fit together?" And then the next question might be, "Did you do anything to train your team or to assess your team or to augment your team to be able to do that project?" And those questions would give you a very clear understanding of what they're thinking about, regardless of whether they're Flinking or Sparking or Hadooping.

ARCHITECTING FOR ANALYTICS

To get a true payoff from your business data, you'll need to make many decisions. There's no one-size-fits-all solution. Before building your analytics system, consider the key factors that can make your project a success.

1. Storing data close to processing can save time and transmission costs. Many companies opt to store all rich data in one location.
2. Real-time analysis creates a different set of demands and may require different tools. For example, you'll need to consider data quantity, formatting, and latency.
3. Access controls should be matched to the sensitivity of the data involved. Moreover, security considerations must protect the data without impeding access for analysts.

ABOUT THE INTERVIEWEE

Bob Rogers

Bob Rogers, PhD, is chief data scientist for big data solutions at Intel, where he applies his experience solving problems with big data and analytics to help Intel build world-class customer solutions. Prior to joining Intel, Bob was cofounder and chief scientist at Apixio, a big data analytics company for healthcare. He believes that accurate understanding of patient conditions, physician behavior, and the characteristics of the healthcare delivery network are foundational to the future of healthcare, and that big data analytics is essential to driving this transformation.

Read Bob's blog on the Intel IT Peer Network.

EXPLORE SOLUTIONS FROM INTEL AND IIA

- Discover how analytics leads to innovation at intel.com/analytics.
- Visit iianalytics.com/services/benchmarking and learn more about IIA's Analytics Maturity Assessment.

experience
what's inside™

Share with Colleagues

0616/AS/MRM/PDF-USA 334272-001

INTERNATIONAL
INSTITUTE FOR
ANALYTICS™