BBBT Podcast Transcript

About the BBBT

The Boulder Business Intelligence Brain Trust, or BBBT, was founded in 2006 by Claudia Imhoff. Its mission is to leverage business intelligence for industry vendors, for its members, who are independent analysts and experts, and for its subscribers, who are practitioners. To accomplish this mission, the BBBT provides a variety of services, centered around vendor presentations.

For more, see: www.bbbt.us.

Vendor:	**Pentaho**
Date recorded:	**November 20, 2015**
Host:	**Claudia Imhoff**, President, BBBT
Guest(s):	**Donna Prlich**, Senior VP of Products and Solutions
	Chuck Yarbrough, Director of Solutions
Run time:	**00:16:55**
Audio link:	Podcast
Transcript:	[See next page]

Claudia Imhoff: Hello, and welcome to this edition of the Boulder BI Brain Trust, or the BBBT. We're a gathering of international consultants, analysts, and experts in business intelligence, who meet with interesting and innovative BI companies here in beautiful Boulder, Colorado. We not only get briefed on the latest news and releases, but we share our ideas with the vendor on where the BI industry is going, and help them with their technological directions and marketing messages. I'm Claudia Imhoff and the BBBT podcasts are produced by my company, Intelligent Solutions.

I'm pleased to introduce my guests today. They are Donna Prlich and Chuck Yarbrough. Donna is the Senior Vice-President of Products and Solutions, and Chuck is the Director of Solutions for Pentaho.

Welcome to you both.

Donna Prlich: Thank you.

Chuck Yarbrough: Thanks, Claudia. I appreciate being here.

CI: All right. Donna, let me start with you. It's been an interesting, six months or so, for Pentaho. Where's Pentaho today? What's been the impact of the Hitachi acquisition on the company?

DP: It's been very exciting, for sure. The Hitachi acquisition was completed in June. It's been the largest big data acquisition to date, which is exciting for us. We're going to be part of Hitachi's social innovation strategy, which is focused around the Internet of Things, and how do we start to connect people, and data, and things. We'll be at the core of that.

We're definitely very much part of their strategic vision for the Internet of Things. We've expanded tremendously. We've added 1,500 sales reps, 50 additional big data specialists. We've got a ton of intellectual expertise, as far as data scientists. There's over 500 now that are at our disposal.

CI: These are Hitachi resources?

DP: Yes. Then, we've also expanded our own headcount by over 30 percent. We've got about 60 percent increase in our consulting staff. Overall, just a really nice way for Pentaho to take what we've done in the past several

years, and being very successful in big data, and now scale with Hitachi behind us.

CI: Excellent. All right. Chuck, let me bring you into the conversation. You talked about the blending of the worlds of big data and the Enterprise Data Warehouse. How did you see that happening?

CY: That's a great question. The data warehouse's been around a long time. We've put a lot of time and effort and thought into preparing data for the analytic process, to be able to understand what's going on in our business. That's what the EDW has done well for us.

The reality is that, in the modern world, we've got new data sources. We got new types. We've got volumes we've never been able to handle before. It's just become much more complex. We know we've got both these worlds. Data is data, whether it's big data, or enterprise data, or whatever.

The reality is, we got to bring them together. We got to be able to blend that data, and blend those worlds together, to give us the answers, the ideas, the insights, that we need to drive our business, to improve the way we deal with customers, to do all those things.

CI: It's been an interesting transition, what you brought forth was the fact that there are in fact, at least two analytic ecosystems now. There is the traditional data warehouse. I would say it's more for production analytics. Then, there's the experimental side.

Let's face it. Streaming sensor data doesn't fit nicely into a data warehouse, so we have to put it somewhere to analyze it. That's the area of the big data lake, or whatever you want to call it, the sandbox experimental area, where we can put this data. Is that the way you see the combination, the blending? It has to be of these two worlds?

CY: I actually see that those worlds are coming together, that they are different. They're driven by different needs, needs of the user, needs of the use-case, and the fact that it's different, different data, different types. Long-term, those things will become tighter.

Today, the challenge is just, how do you bring that data together to make sense? Because what's in your big data, you mentioned sensors.

Sensor's a great example of big data. It doesn't fit well into a data warehouse. Sometimes, sensor data is only valid for a moment. You're not looking at long periods of time…a data warehouse has a time dimension typically. But, without the context that might come from a data warehouse, that sensor data may not have any value.

How do we make that happen? How do we blend those worlds? You got to have some technology to do that. That's where Pentaho plays really well. Customers are using datasets together, to make good decisions, drive business, predict the future.

CI: I get it. Donna, let me go back to you. I do want to dive in a little bit deeper into this idea of blending. You put up a slide that was the analytics data pipeline ecosystem. I found it to be quite interesting.

Again, in more detail a little bit about this blending of big data with the Enterprise Data Warehouse… how do you see that helping companies with this duality, if you will?

DP: There's a couple of areas where it makes a lot of sense. If you think about Chuck talking about a typical Enterprise Data Warehouse environment, and then you've got a big data environment. In some ways, it's a one-dimensional view. The fact of the matter is that, data is flowing throughout those different places, whether it's in the data warehouse, or in Hadoop, or in some other type of data store.

We think about it as data in motion. If you think about data's constantly moving, what do you need to do to manage it? We have to think about how it flows through a pipeline. What we found is, this concept of an analytic pipeline helps our customers to think about the data engineering, which we think of as traditional ETL. But it's beyond that. It's ingesting data. It's cleansing it. It's all of those things.

Then, that is associated with somebody who is probably a developer. Then, you've got this center area of the pipeline that's more focused on

things, like profiling, refining. The folks who are really digging in and trying to figure out what are the best slices of that data.

Then, you've got the Endgame analytics. We see that as, that data can flow out to somebody who's embedding analytics in an application that's out on the ship at sea, or it can be somebody who's sitting and wants to be able to use R to refine datasets and look for interesting patterns.

We see that flow. You've got to think of that whole thing. We think about the bottom of that pipeline that's really essential to make sure that it's being managed.

We've put a lot of things into our latest release that focused on things like lineage, monitoring, security, because that data's flowing, but at the end of the day, it still needs to subscribe to a lot of the things we've done in the past, and thinking about how we manage and make sure that data's accurate, and trusted, and all of those things.

CI: Very critical I think. All right. Chuck, let me go back to you. What do you see as the drivers for big data? I know that you said that big data eventually will become just data. I applaud you for that. You're right. But, there are drivers for big data right now, aren't' there?

CY: Absolutely. It's interesting, because when we sat back in the early days of the big data movement, and there's a lot of hype, a lot of people talking about what it could do. There's also people saying, "Hey, we're not getting any value. We're just spinning our wheels." At Pentaho, we had a lot of customers, who were, frankly, they were early adopters of big data.

They were getting real value. We sat back as a team, talked to our customers, looked at what they are doing. We identified three real key drivers. If you didn't have one of those drivers, it was probably a science project. That was the reality.

Those drivers were as simple as 1) Did it drive incremental revenue?

Were you able to predict behavior and understand what customers needed, wanted, or were doing?

Did it improve operational effectiveness?

Can we reduce financial risk and manage fraud?

Can we reduce data warehouse cost?

Things like that.

Then, the third was improving the customer experience.

In today's world, the "age of the customer" as some people call it, are we driving the personalization to the level that we need to? If you look at those organizations that are doing that, not in a creepy way, but in a good way, they're really executing well and reaping the benefit of the investments in big data.

CI: They certainly are. The one that comes to mind is something like situational one-to-one. It's no longer just, "I understand who you are, but I understand who you are in the situation you are in right now." We're getting a real, fine point on that. How does Pentaho then help companies with these drivers?

CY: I talked about how we sat back and looked at what our customers were doing. We identified what their drivers were. We also identified a number of use cases, repeatable use cases. We call those our blueprints. Think of them as recipes for success. Ways of helping customers understand how they can leverage the technology, both Pentaho and other technologies to do something.

For example, data warehouse optimization. A lot of people are spending a lot of money on a data warehouse, data warehouses that are expanding and getting too big, maybe too unwieldy. There's a pattern that allows them to fairly easily implement big data technologies, to reduce their costs, and at the same time, improve what they're delivering, meet their SLAs, deliver to their customers better, easier, cheaper.

Customer 360 is another blueprint that we do a lot of, again, going back to the customer experience. It was those blueprints that enabled our

customers to adapt the technology in an easy way and fit it into their organization.

CI: I like the blueprints, because they're reusable components. You figured out this is a pattern. I might as well bundle it up, and go ahead and say, "Look. This is the pattern you're following. We've got something for you." Right?

CY: That's absolutely right. Any of us that have been around for a while... we remember the early days of data warehousing and those patterns, those dimensional models. Those things didn't just happen. It took time. It took effort. In a way, the big data world is similar.

We have the technology, but we have to apply really intelligent approaches to enable us to deliver the right solution, the right outcome, that the customer needs.

CI: I agree. Donna, back to you. Let's go into the contrasting approaches to your analytic data pipeline, if you don't mind. What's the contrast here?

DP: We talked about data in motion. It doesn't necessarily only sit in one place. You've got a data warehouse that is running all of your day-to-day operations, for instance. You've got a data lake, where maybe you're housing multiple varieties and sources of data, and you're doing some kind of exploration of that data. Or maybe you want to take that data, and blend it with data in a data warehouse.

A data refinery, we often see that's one of the blueprints that Chuck mentioned that customers will put data into Hadoop for processing. What they want to do is, they want to be able to deliver slices of that data to different users for different business purposes. The real value of having those contrasting approaches is, data's going to flow into anyone or two or three possibly, depending on what a customer needs.

It's important to think of the data as a pipeline that feeds those different places where data lives. If you're looking at it that way, you're going to have a lot easier time bringing those two worlds together. We talked about the data warehouse and the data lake. We found with our

customers thinking about the blueprints, one of the comments you made Chuck was packaging that up and recommending it to customers.

The other thing we've seen is oftentimes a data warehouse offload, where somebody simply wants to take data out of a data warehouse, purely for cost reasons, "If we take this out, we save extra number of dollars per year, a good idea, put it in Hadoop, it's less expensive…" we'll often see that evolve into a data refinery architecture with Pentaho, because now, they have access to data that they didn't before, because it sat somewhere that was hard to get to.

They realize the blending capabilities. Then from there, they might start to refine data out to different types of users.

Those contrasting approaches are all valid… is really the point. Overtime, it's important to plan to either have them, or maybe you're going to change your approach, and think about all of your data in that context, in order to be successful.

CI: I like that. I want to stay with you, Donna. If you don't mind, tell me about FINRA. It's an incredible case study. It's an important one. If you don't mind?

DP: Yeah. It's really fascinating. They're a financial services, regulatory agency. They process, on a really heavy day, 75 billion transactions per day. They have this data that comes in. They're really looking for fraud. They're looking for people who are trying to play the system. They have a team of analysts that needs to be able to get at that data in an efficient way.

What we've helped them to do is to take the data in Hadoop, refine slices of that data out to the analysts, based on what they're looking for… Certain stock symbol at a certain time was being traded by certain brokers… And very specific information, automatically run a transformation that blends all of that data, delivers it to these analysts on demand. Then, they're able to match more easily and efficiently detect fraud.

It's a great example of the two worlds coming together, empowering the user to get access to that data, but in a very well-governed fashion. As you can imagine, somebody like FINRA, that data's got to be governed.

They have to know where it came from, and make it repeatable, if they need to go back and regenerate that dataset. They've optimized the efficiency of those analysts in an amazing way, by giving them the power to get the data themselves.

CI: That's a cool case study. I have to admit, it's a really cool case study.

All right. Chuck, back to you. You've got Pentaho version 6.0 out already. If you don't mind, just spend a minute or so and tell me about the key features of 6.0.

CY: OK. Yeah. Released about a month ago at Pentaho World. 6.0 is a big release for us, pretty foundational. Some of the key features for us were things like this idea of taking a complex transformation, and instead of taking that data and transforming it, conforming it, and then putting it somewhere, persisting it in a table or in a data lake, rather, making that data available as a data service.

So Pentaho services is a key feature in 6.0... Something that our customers have been wanting to do, with all the right pushdown optimization and things to make high performance big data. You could think of it as a virtualized dataset, where that transformation just becomes available instantaneously.

A lot of our effort in this release was also around the enterprise manageability, extending capabilities around data lineage, making it easier to not only track lineage, but to bring lineage from Pentaho into something else as well and share that lineage. Life cycle management security is always a big topic. Our customers are feeling the pain, and know what they need to do. Those are the key features.

We also included a lot around the automation of that pipeline. Once data is processed, prepared, and ready for the analytic process, to automate the delivery of that, whether it's to user as an analytic model

and a visualization, or to run that through an advanced algorithm through Weka or R.

CI: All right. A lot of stuff there, I have to admit. Unfortunately, we're out of time. That's it for this edition of the BBBT Podcast. Again, I'm Claudia Imhoff. It's always such a pleasure to speak with Donna Prlich and Chuck Yarbrough of Pentaho today. Thanks so much for speaking with me.

DP: Thanks, Claudia.

CY: Thanks, Claudia.

CI: I hope you enjoyed today's podcast. You'll find more podcasts from other vendors at our web site www.bbbt.us. If you want to read more about today's session, please search for our hash tag on Twitter. That's #BBBT. And please join me again for another interview. Good bye and good business!