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:	**Talend**
Date recorded:	**July 26, 2013**
Host:	**Claudia Imhoff**, President, BBBT
Guest(s):	**Yves de Montcheuil**, VP, Marketing
Run time:	**00:11:23**
Audio link:	**Talend Podcast**
Transcript:	[See next page]
Transcribed By:	CastingWords

Claudia Imhoff: Welcome to this special edition of the Boulder BI Brain Trust podcast. I'm Claudia Imhoff and I'm here in Grants Pass, Oregon, at the Humphrey Strategic Communications annual Pacific Northwest BI Summit. I'm pleased to have the opportunity to interview each of the vendors attending this event.

CI: With me is Yves de Montcheuil. He's the vice president of marketing at Talend. Welcome, Yves.

Yves de Montcheuil: Well, thanks, Claudia. It's good to be here.

CI: We've discussed a number of trends in the BI Summit this year. One of them, of course, were all the trends around Big Data. I'd like to ask your opinion about some of these trends. What do you think?

YdM: Big Data is certainly a very interesting topic. It's a little bit over-hyped, but it's becoming less and less over-hyped. As we have seen in the past, I would say, 12 to 24 months, the market has evolved from experimenting, sand-boxing around big data into starting to actually get value out of it, deploy real projects. We are seeing customers who are actually being successful with their big-data projects, who are telling us how it's actually working, what's working, what's not working.

CI: I agree. What I've heard are some very good case studies -- down-to-earth, realistic, good benefits around Big Data. I'm very happy with that trend. I think the hype over, "Oh, it's only social media," for example, that's going away. People are realizing that it's coming from all over the place, isn't it?

YdM: It's coming from all over the places. Social media is certainly one angle of Big Data, but so is, frankly, your traditional transactional data coming from your ERP and your CRM. Then you've got another type of data that belongs to the Big Data project. It's what we like to call the dark data.

CI: Oh, dark data.

YdM: Dark data is data that is produced by systems but that are exploited, it's stored somewhere -- log files, in servers, on desktops, whatever -- but nobody uses it. Then you have even further than dark data, the exhaust

data -- the data that goes in fume, the data that your car produces, several terabytes per day of driving -- that's, frankly, only used by the dealer when there is an anomaly, to fix it. Think what the benefit could be to actually analyze this data, like the aircraft manufacturers do with the data that's coming out of the exhaust of their airplanes, for predictive maintenance, for your analysis of the driving styles for better longevity of the cars.

Bring all those data together. Don't leave any data behind.

CI: Speaking of that, I understand you can give me some good tips and techniques on handling Big Data. Why don't we start with some of those?

YdM: One thing we are saying is that big data is not only about being big. You should forget about volume. Volume is not relevant. Big Data can be large. It can be small. It can be fast. It can be slow data. In some industries, organizations will deal with gigantic amounts of data. It's just the nature of the way they do business. In some organizations, a couple million records is already a very large dataset.

CI: Do you see then that it's more the complexity of the data, the complexity of the queries, rather than just the volume of the data?

YdM: It's the complexity. It's the variety. It's making sure you go after a comprehensive set of data so you know that no data left behind is obviously extremely important there. It doesn't really matter whether your Hadoop cluster is going to start with 12 nodes or 500 nodes. Don't be complex by the size of your Hadoop cluster. It's really not what matters. It's the use cases and the value you will be able to derive from your data that really counts.

CI: What other tips do you have for me?

YdM: One thing we are seeing is a trend to not get every single piece of data into Hadoop. Hadoop shouldn't be a big data dump. Some data is already in the data warehouse, and the data warehouse is a very good place for some of this data to reside. There is no point in duplicating it. Link it to Hadoop so that it can be used by the same type of analytics processes. Some of the data will be better off residing in the source

systems. In the MDM world, this concept of distributed MDM, reference MDM, the logical data warehouse model where you don't actually consolidate everything in the data warehouse. While you can do something with Big Data, you can have a big logical data system.

CI: In other words, it does not all either have to be in the data warehouse or in the Hadoop system, or in the non-relational system. The key though is that we have to have access to both of these, free access so that we can combine it perhaps, or twist it or turn it and look at it in different ways.

YdM: That's where integration obviously comes into play. You need to provide the link between those systems. You need to provide the piping and the plumbing to move the data between those systems. Even beyond exchanging data between the different constituencies of the logical big data system, Big Data must be part of the overall IT infrastructure. Don't treat Big Data as an isolated island. You don't want to cast away your Big Data in this island in the middle of the ocean, like poor Tom Hanks in the movie.

CI: No, poor Tom Hanks!

YdM: The sandbox is fine for a proof of concept, the Hadoop cluster in the closet not linked to anything else, or just that you provision with your credit card in the cloud. That's fine for trying it out, for figuring out what you can do, but then, as the Hadoop cluster starts to become a source of trusted information for other applications, something that you will connect your analytics systems to, it must become part of real IT.

Real IT that's managed by IT, that's supported, that's maintained, and that's able to exchange information in a better form than flat files, that get emailed or brought on a USB key because the cluster is in a closet, and it's isolated from the rest for security reasons.

CI: What I've heard is that a lot of companies are using these non-relational environments, more as an experimental area, "Let me bring the data in, let me analyze it, let me find what's of value." There's a lot of noise in that data that is simply not useful to the organization. Maybe once you've found the information that is actually of value, maybe you might want to even move that into your data warehouse.

YdM: But it's an ongoing process. You don't find that information only once. Of course, initially, it's an iterative process. It's like in predictive analytics. You first bend the model. You figure out what the model is, but it's not a one-time analysis kind of thing. Again, it's an ongoing process. Once you have understood what value you can derive from your big data, then it must become part of the overall application. Not only IT, but application infrastructure. That's another very important angle. Big Data is not only a data platform. It's becoming an infrastructure. It's becoming an application platform.

The end goal of big data is not to store and process this data. It's to provide services to other applications that will be fed by this Big Data, by those nuggets that you have found that will allow you to provide more value faster. Think the Amazon recommendation engine. It was developed way before Hadoop was invented.

CI: Exactly, and worked pretty well.

YdM: It worked pretty well, because Amazon invested a lot of money, a lot of development efforts, and probably a lot of processing power for that. If they were to redo it today, my guess is that they would redo it in Hadoop. What's available now is the ability for any second tier or third tier website to actually build a similar recommendation engine. Recommendation engine that's not a one-shot thing that you then provide the results to the buyers, but it's really an ongoing process that is going to feed the website in real time as people explore products, articles, et cetera.

That's a very good example of Hadoop becoming a Big Data engine, or becoming an application that is actually fitting the bottom line of the enterprise.

CI: Of course, one of your last tips has got to be "Watch your costs with Big Data."

YdM: Big Data has a very interesting angle, in that it was born as open source, it continues to live as open source. I actually believe that because Big Data technologies are open source, that's why innovation happens so quickly in the Big Data space. Now, brings of course another very interesting angle, which is that the technology is either free or at least extremely affordable.

We all know that "open source" is not synonymous with "free," but there is enough competition out there, and there is enough opportunities to try before you buy, and all that good stuff, that it's really driving the prices down, which is also a big factor in making the technology affordable for all types of organizations, not just the Amazons and eBays and Citibanks of the world.

CI: Right. Very exciting time for the open source vendors like Talend, right?

YdM: It is a very exciting time for the entire open source community.

CI: Of course, it is. Thanks so much, Yves. My guest is Yves de Montcheuil. He is the vice president of marketing at Talend. Thanks so much, Yves.

YdM: Thanks so much, Claudia. That was great.

CI: Thank you for listening to this special edition of the BBBT podcast, and thanks to Scott Humphrey for giving me this opportunity and for hosting the Pacific Northwest BI Summit.