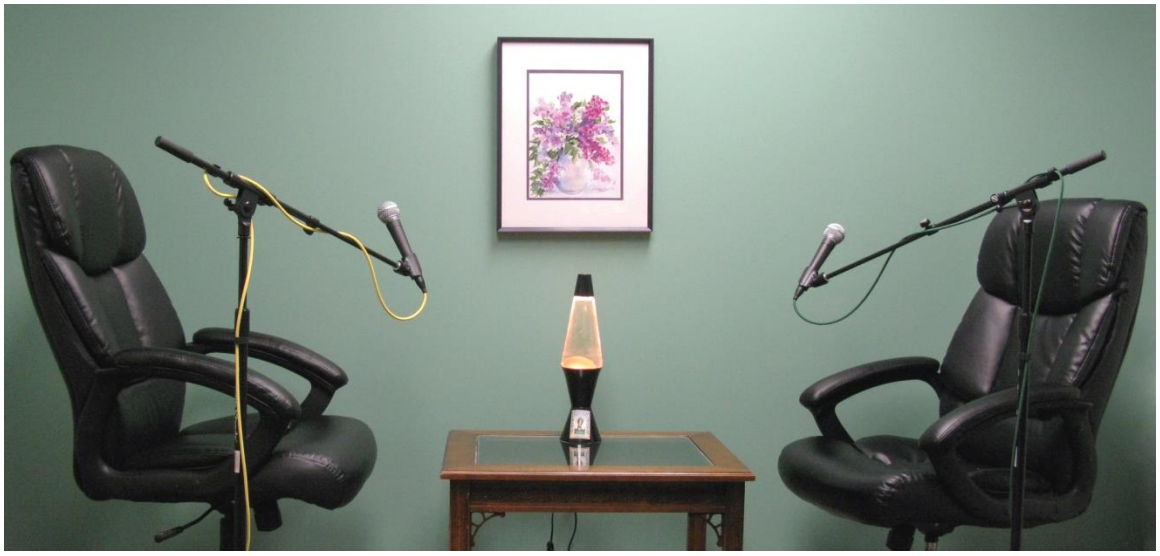
BBBT Podcast Transcript



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:	**Hortonworks**
Date recorded:	**February 14, 2014**
Host:	**Claudia Imhoff**, President, BBBT
Guest(s):	**Jim Walker**, Director of Product Marketing
Run time:	**00:16:57**
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 guest today. He's Jim Walker. Jim is the director of product marketing for Hortonworks and a charming young man, very entertaining. This should be an excellent podcast. Welcome, Jim.

Jim Walker: Thank you, Claudia. I don't know how young I am, but I'll take it. Thanks for having me.

CI: You're welcome. Let's start off with Hortonworks' vision. It was all about the modern data architecture. Why don't you just quickly give us an overview of that architecture?

JW: If you look at the front page of our website, it says, "We do Hadoop." Our vision is really centered around making sure that Hadoop is an enterprise viable data platform. Making sure that it's used by everybody. Making sure that it's a plus one in the data center.

We don't believe that Hadoop is going to obviate anything in the data center. We believe it's not going to block out the sun. We're really more about making sure that the rest of the ecosystem and the partners you rely on today, that they can rely on Hortonworks.

If Hadoop is going to be a platform for the future that provides us great value in terms of linear scale storage and compute, the promise of that is fantastic. The challenges, do you have the resources and the skills to actually take advantage of that? Do we have the wherewithal to insert something in my data architecture?

We want to make that as simple as possible. Really, the modern data architecture is really about relieving some of those pressures, and it's really working within the ecosystem that's already existing and you are already

engaged with, to make sure that Hadoop can augment and help provide further value to organizations.

CI: Let's drill into that just a little bit further, because there are some vendors making noises about how Hadoop can replace the enterprise data warehouse. It's a rip and replace philosophy sort of thing.

From what you've said today, that's not your philosophy at Hortonworks. Why don't you expand on why not, first of all, and secondly, what is your philosophy in this situation?

JW: Boy, Claudia, you really like to scratch that one with me, right? Let's be pragmatic. Let's take all the aspirational marketing aside and let's just be truthful. The Enterprise Data Warehouse has been around for how long? 30 years?

CI: Easily.

JW: 25 years? The functionality that we find in this awesome system is just that, awesome. It's mature. Organizations have been relying on the Enterprise Data Warehouse for years. It does provide value. There's much more to the enterprise data warehouse than passing data or asking explicit questions.

There is iterations over data in the warehouse. It does provide multiple different functions. There is a lot of technology in terms of how many concurrent users can access Enterprise Data Warehouse.

Earlier today in our session, Neil Raden was asking about that. It's a huge piece of it. By no means do we want to replace the Enterprise Date Warehouse, we want to augment. We want to make sure that sets of data that we once thought were not analyzable, I guess, if that's the word, can be analyzed. Can they be used in the context of existing data that's in the EDW?

That really requires a two way street. Can I take data from the EDW, bring it into Hadoop, combine that, munge that with data like clickstream, and sentiment, and server, and geolocation, these larger data sources, get some detail out of that, then feed it back, and maybe augment the data

that I just pulled in, in my EDW, with something that means value to an organization. Our customers are delivering this. Our customers are doing this. But more importantly, our partners feel that this is very important as well.

Again, I guess all my questions might come back to really a lot of our partnerships, in that we want to work within the ecosystem. If you look at what we're doing with the Microsofts of the world, SAS, ASP, Teradata. I mean, can you get a larger EDW footprint on the planet? They're working with us, and I think that's an indicator.

It's often confusing for people to understand what's happening in the Hadoop market. If you really want to know, maybe take a macro level picture of what's going on and look at what the big venders are doing and how that...does ETW work with Hadoop? Let's go out and ask Teradata and some of the other leaders out there. So, it definitely works together.

CI: Let's drill in a little bit more here. This will lead into it. You talked about something called a "Data Lake," and it grew out of the need to offload, if you will, some of the more difficult things that perhaps the Data Warehouse shouldn't be dealing with. Why don't you describe that a little bit and, more importantly, see if you can define this Data Lake for us.

JW: The Data Lake's a little bit of an overloaded term. I think sometimes you call it Data Reservoir, Data Lake, Data Pond, swimming pool,

CI: Data Repository, Data Refinery,

JW: Repository, yes. There is a whole bunch of things. Really, the Data Lake is, I mean, a small percentage of where our customers are. I think at the grand maturity level in Hadoop, people want to store a lot of data in one single place and to interact with it, simultaneously in multiple different ways. That means looking at data from many different angles. By angles, I mean providing access to data via different engines, if you will. What's the most common way of interacting with data? The English of the data world is SQL. I have to have a SQL engine on this.

Can I also store all my data and interact with it via search? Can I have some sort of in memory, things like Spark and these other type of

technologies, where I can have a shadow when I ask a question? These sort of things.

A Data Lake is really providing a grand repository where it makes sense to store all the data, and then ask various different questions on it. Again, it's kind of the end of the maturity game for one who's going to start with Hadoop. Most companies start with a net new analytic application, where they're getting a very discrete value out of it. I think that's the pragmatic approach. Like we've seen in the past, as you grow more and more data in a growing system, man, it might be interesting to look at the combination of these sets of data, and that's where I think people start to go down that path of, "Is it a data lake?," whatever you call it. I call it one really great, huge repository to ask some really cool questions in many different ways.

CI: What I think is interesting is that it's a way for a company to bring in big data, this unusual data, geospatial, whatever it is, sensor data, whatever. Things that don't quite fit nicely into an EDW and at least bring them into an environment where you then have the opportunity to analyze it, to discover it, to figure out which pieces of information are actually of value. Maybe the rest of it has no value; why would I put that into my data warehouse?

JW: I can't agree any more, yeah.

CI: Then you have the decision to make of, do I leave it in the lake or is this something that I actually ought to properly put through ETL grist mill and data quality and move it into my EDW?

JW: Yeah, and quite honestly, Claudia, I would replace in your last sentence "Hadoop" with "lake". Honestly, people are using Hadoop to chunk through sensor data. Say you're a building maintenance company and you have elevators and you have chillers in the basement of the place, air conditioners, if you will. Tons of different sensor data in the building; how do I actually process the data so I can do predictive analytics across it and say, "I need to repair over replace"?

It's those discreet applications that people are using Hadoop for. Does it turn into a data lake eventually? Sure, probably. That's when you start to do things beyond just, say, sensor data, and you add in clickstream data or

you add in geolocation data. Wouldn't it be great if I could do geolocation data with sensor data, say in the same building company?

I'm identifying when my equipment is failing, but you know what? I also know in my ERP system that this particular repair guy is out there working on those things. I could look at the geolcation data around where that guy is going and how he gets to the building.

There's a lot of different ways that we can munge data together to actually accrue new insights into it. I think that's where the value of starting to move different types of data. Ultimately, that's all going to get boiled down and used in the context of, say, an EDW or your BI tool, because you've got to visualize it, you've got to…It's about using it to chunk through and then use the data.

CI: It's almost as though once it becomes a routine query, that's the point where we want to move it into the EDW, because why would you waste all this wonderful experimental capabilities for a plow in a field? Not that the warehouse is a plow horse, but you get the idea.

JW: It's like science. I do a bunch of exploration and I run a lot of iterations and I look at this chemical in a different way and how it mixes with this and then all of a sudden, I've created a formula...

CI: Then you can do production.

JW: I've created a formula for marshmallow fluff. Let's put it in production; let's get it out there.

CI: Nutter Butter, baby.

JW: Exactly.

CI: Let's get to Hortonworks, in particular, then. The product, if you will, that you guys offer is really putting all of the little pieces together. That's sort of a simplistic view, my simplistic view, but there are so many little pieces to making Hadoop a full-fledged environment.

It seems to me that the best way to describe what Hortonworks does is you guys make it work together. You make everything just fit and continue

seamlessly. Somebody could do that, obviously, it's all open source; they could do it, but it would take them years to put all these things together, and a team of...how many people do you have? Is that a fair assessment or is it a little more than that?

JW: It's a little bit of an overstatement, in that yeah, sure, it's a product, but that's free.

CI: Product does not mean you have to sell it.

JW: Exactly, right. We create the Hortonworks data platform, HDPs, as it's lovingly called, internally. Yeah, we do pull all the various different pieces, across all the various different releases, within these Apache projects. There's Apache Hadoop, Hive, Pig, Hbase, Zookeeper, Scoop, Floom. It's like your six children, you always forget one, right?

There're all these different projects. Yeah, we do apply our...

CI: It sounds like the seven dwarves, to tell you the truth.

JW: Except Grumpy.

CI: No Grumpy.

JW: We do pull all that in, and we apply our expertise that we've learned, really since 2006. We apply the appropriate enterprise rigor to our distributions, so that our customers can rely on it. That's absolutely imperative to us.

We want Hadoop to be everywhere. If Hadoop is going to be able to work, people have to rely on it. I can't have some untested software put into production, and then I'm going to fix it. It doesn't work that way.

Yeah, we do apply rigor over that. The main product we sell is support and services around it. Because we can do that, we can provide some of the best support and services around what we do.

CI: Well quickly, one of the big changes coming up, or actually in place now, is YARN. That's changed the platform pretty significantly. Can you give me a brief overview, in a minute or two, of what's going on?

JW: Sure. YARN changed the platform. YARN is a game changer in the way that we think about data. Honestly, I like to think about it as a data operating system, where you have access to this linear scale compute and storage, and you can start to store tons of data at a much more cost effective manner. Then start to interact with the data, as if you had an operating system for it.

That's really what YARN provides. It provides all the basic frameworks that you would need, across reliability and security, and all these different things. Yet still being able to do these things that are interacting with data, in multiple ways, simultaneously.

Hadoop is really gone from being, originally, a batch system. 2008, if I had a couple terabytes of data, a petabyte of data, and I wanted to run a query across that, "Great! Oh, my God, I'm able to do that in a day, awesome! I just wasn't able to do that before." People want to interact with data in seconds, having real time interactions with that, yet still have the batch stuff and also have streaming and being able to pick a fence off of data within Hadoop.

These things become very important. "Can I do this all in a single place, in a single Hadoop cluster?" That's what YARN really enables. There's a bunch of LAN architectures that are being built, that say, "This is how we do this in a particular business." Oh, man, the business applications and use cases that can be built up around having one single set of data as far reaching as we can possible get, and then using different engines to look at that data in multiple different ways. It scales tremendously, and it's going to open up a lot of us to a new way of looking at data.

CI: Yeah, I think I agree with you. Now you mentioned partners. *Everybody* has partners.

JW: Yup.

CI: Some people have their logos on their website, and all it really means is that they have a connector to their database, or a user interface, or whatever it is. Something like that. Many of your partners are actually major, relational DBMS vendors. You mentioned them, Microsoft, IBM, Teradata, and so forth, SAS, and on it goes, SAP.

Are these just surface relationships? Is it really just a logo on a website, or do you have a deeper, some real skin in the game with these guys.

JW: Yeah, it's definitely skin in the game. The way we look at partnerships, we don't look at them as opportunistic from a market point of view. Absolutely, we're opportunistic, in terms of getting to customers and these sorts of things, but we really look at it as a way of extending Hadoop and making sure that Hadoop is going to be used across all the data centers in the world.

If you look at what we do with people like Microsoft, and integration with SQL Server, Parallel Data Warehouse, and Excel, oh, my gosh is Excel not everywhere? If you look at what we're doing with SAS, and putting the SAS analytics engine directly in YARN. If you look at the conduit that we're putting between SAP HANA and Hortonworks. If you look at what we're doing with J2E at Red Hat and Hadoop. And if you look at what we're doing with Teradata, and the work we've done with them with SQL H.

All of these relationships that I just spoke through, all point back to an engineering relationship. There is a piece of technology that is actually being developed in conjunction with these partners, that makes their solutions work very well with Hadoop.

Is this open for everybody? Yes, we do all of our work in the open, so ultimately, all of Hadoop, I don't care which vendor, and all of the different pieces, and whoever it is, will ultimately benefit from this. For us, on the other end of this, is our partners know that we are best to support that particular relationship.

Can everybody use it? Can everybody benefit from it? Absolutely. It's really core to our philosophy. We've hit all the points really around our philosophy, it's make sure Hadoop provides all the functions it needs to, as a platform.

Number two, apply rigor so that people can depend on it and rely on it. Really, three, which is really since day one, is make sure the ecosystem can integrate and work with Hadoop. That really comes down to it, that's the three things we focus on.

CI: Excellent. All right, you've got 30 seconds. Tell me what the future holds for Hadoop.

JW: 30 seconds to describe, really, the future of Hadoop? Man, I could go on, on this one, for quite some time. We're just really at this genesis of this transformation. As we've spoken about, Hadoop is a plus one in the data center. Tighter integration with all the tools and all the skills that we have is going to be really critical.

You know what we're going to go through? More of this disillusionment, if you will, with Hadoop. People, they doubt, because there's a lot of hype. What we're going to see is a lot of reality come out this year. We're going to see a lot of vertical applications being built on Hadoop.

When that happens, we're going to truly have crossed the chasm, if you will, to borrow from Geoffrey Moore, and we'll see it become this mainstream technology. It's mainstream today in certain groups. But we're going to see broad adoption of Hadoop over the next couple of years.

It is here to stay. I don't think there's any way to argue that there's not an organization on the planet that's either considering or thinking about evaluating Hadoop in some form or fashion. It's broad adoption.

CI: Not only is the elephant in the room, but people are talking about it.

JW: Absolutely.

CI: All right. That's it for this edition of the BBBT Podcast. Again, I'm Claudia Imhoff. It's been a great pleasure, and really, really nice to speak with Jim Walker of Hortonworks, today. Thanks again, Jim.

JW: Thank you for having me, 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!