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:	**MapR**
Date recorded:	**January 23, 2015**
Host:	**Claudia Imhoff**, President, BBBT
Guest(s):	**Steve Wooledge**, Vice President Marketing
	Tomer Shiran, Vice President Management
Run time:	**00:20:06**
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 Steve Wooledge and Tomer Shiran. Steve is the Vice President of Product Marketing, and the Tomer is the Vice President of Product Management for MapR. Welcome to you both.

Steve Wooledge: Thank you. Glad to be here.

Tomer Shiran: Great to be here.

CI: Nice to have you both. It was a most interesting session. Let me just start off that way.

I hope people do watch the video, because it really was an excellent session, a lot of good information from you, a lot of feedback from the members, but I thought a really good one.

Let's start off with a question to you, Steve. Why don't you tell me a little bit about MapR for people who may not know what the company's all about?

SW: Sure. MapR is a software company based in San Jose, California. We've been in business since 2009. We were one of the big three, if you will, from a Hadoop perspective, so we're a Hadoop distribution software company.

Our focus is really on production success for Hadoop with large scale and medium scale customers to get the most value out of their data.

CI: 700 customers, right? More than? All over the world?

SW: Yes, over 700 customers. There are about 300 employees now in the company, expanding globally in every continent across the world and country.

CI: And a stellar set of partners.

SW: We've got about 300 partners today, a lot in the data management, BI analytic spaces you would imagine, the Tableaus, MicroStrategy of the world. We are resold through a number of partners, including Cisco, Teradata and HP. Really, it's based on customer demand, so if a customer has needs for a certain certification with a certain technology, we'll go ahead and do that. It continues to grow out as our customer base expands.

CI: You guys are on fire. Let's talk about your customers a little bit. You gave us a number of customer case studies that were really interesting and kind of different ones. Why don't you give me two or three of those?

SW: Some of the interesting ones are customers that have been using MapR and Hadoop for a long time. They power many different types of use cases, so if you take one in financial services as an example, they've created a data link or a data hub where they're powering not only fraud investigation for their credit card transactions, but they can also power the real-time fraud detection transactions that need to happen in real-time.

They're also powering other applications such as recommendation engines, so based on the information they have about you as an individual and the vendors or merchants that you shop with, they can provide more relevant offers and discounts through some of their merchant partners back to you as a consumer. And they can do that through one system on one pool of data, because of the scale and the performance that they get with MapR to be able to support not only the analytical workloads, but also some of the transactional workloads.

CI: The other one that I thought was interesting was the retail one, especially at the very end when you talked about someone who is checking out of the store. The point-of-sale system, I'm assuming, can actually do some kind of a scan of the Internet to see if there's anything else priced cheaper than what they're paying?

SW: Yeah, that's right. They've done price comparisons for a long time to make sure they always had the best price in store, but they were able to speed up the time it took to do that and make it real-time, so as you're checking out, if there's a better deal online from any other vendor, they'll match that

and give it to you as a credit to come back and shop again and use that credit that you got back.

CI: Fascinating, absolutely fascinating. Let's get into MapR then a little bit. When you say you support Hadoop, that means a lot of different things or your Hadoop distributor. It means a lot of different things. What's involved in your support of the Hadoop platform? What is your architecture?

SW: We are a Hadoop distribution, so we package together many of the open source projects, some of which we developed, some we contribute to, some that we take back from the community, but we run it on our data platform, the MapR data platform.

That's where a lot of our engineering effort has gone, to make all the projects and other software, both within Hadoop and other open source technologies, make them more reliable, more performance, so that the customer can get more value from all the community innovation that's happening.

I don't know, Tomer, if you'd like to add more to that from an architectural perspective.

TS: I think the design center for MapR is really to make Hadoop look a lot more like other things that companies run in their data centers, in terms of the reliability and the business continuity capabilities.

If you think of databases or data warehouses or enterprise storage, most companies will have strong requirements around high availability, data protection and disaster recovery. Those are things that MapR uniquely brings to the table in the Hadoop world.

If companies want to run Hadoop in production, then typically they'd want those kinds of characteristics. MapR's the only distribution that can actually provide those.

CI: It's fascinating to me. You did have a very nice comparison, if you will, with the other two big distributors of Hadoop.

That's my next question. Why would I pick MapR over these other two? What are the differentiators there? That's to both of you because I think you both have a contribution to that.

SW: I think Tomer said it well in terms of the reliability and the availability in disaster recovery. Those are the core things that we provide. That's the number one reason why customers will choose us. That was through a survey that we did directly with our customers.

The second thing is performance, just getting more throughput, value processing, whatever you want to call it, though the system.

The third one is the NFS, the network file system API, that we support the ability to get data in and out of the system much more easily. It sounds really low level and trivial, but when you're talking about literally petabytes of data, that ability to treat Hadoop like another system within your architecture with a standard interface is hugely valuable.

Those are the top three.

The other one is multi-tenancy. With all the examples I've given, with another distribution, you typically need to create separate clusters for these different applications. It's like the data mart spread that we saw in the data processing industry, where now you've got data in all these different silos. Well, if you have a true big data platform that can support multiple workloads and applications and users and segment it by geography or workgroup or what have you, you're going to get a lot more economies of scale, but also a lot less data movement and all the things that we've seen in the industry for a long time.

That's a really big deal for customers that are going into full-scale production with Hadoop to have that native multi-tenancy and everything in the platform.

TS: Steve talked a lot about the operational advantages, which would certainly, in any organization especially the IT department, really value.

There's another big difference between MapR and other distributions and that's the fact that it's the only real-time Hadoop distribution that enables

people to analyze and process the latest data and run operational applications.

I'll give you some examples of capabilities in MapR that enable what we call the "as it happens" business, the ability to respond to things that are happening now.

It starts with the ability to ingest data into the platform in real-time as opposed to batch uploading it once an hour or once a day. Data can be streamed directly into a MapR cluster and you can do analysis on the event that happened one second ago. That enables, of course, real-time map targeting or real-time promotions or real-time security and detection of security events.

Another aspect of that is with our Apache Drill capability, that ability to then analyze that data in real time because if the data landed in the cluster in real-time, but you then have to wait on somebody else to define a schema, for example, and prepare that data for analysis, you're not getting that real-time value either.

The third thing is the ability to run operational applications directly on the platform. With MapR, we focused not only on the analytics and the processing of data, but also the ability to run a live database application directly on MapR, as opposed to having a separate database and analytics environment—you can actually have those things in a single platform, and then do operational analytics and analyze the live, real-time data of that application.

CI: Such a critical thing, especially for risk and fraud and those really sensitive...Like you said, security breaches, all of those really sensitive areas that cost companies billions.

Excellent!

Let me get into a little more detail here. A lot of confusion, I think, in my community, the technical community, about where and when should people deploy these non-relational data stores versus the relational ones.

The first question that I always get asked is, "Well, is non-relational going to replace the relational world? Should I just chuck my data warehouse and go down the road of the non-relationals?"

Let me ask you that question and see what your response is.

TS: I think that's an interesting question. First of all, the reality is that mainframes are still around, so nothing is going away or disappearing overnight.

The relational database…you think about when it was invented, that was around 1970. The requirements at that time were very different, or the environment at that time was very different than it is today.

Some examples of that, data was measured in megabytes, some years later in gigabytes. The data was always very structured and the application development profiles, the way in which applications were built, was very much planned. You'd spend a lot of time planning the application. Then you'd go build it for a year. Then, you'd ship it to customers in a box.

Today, when we look at our customers, they're building Web and mobile apps and they release a new version every day, sometimes multiple times a day. The volumes of data are very different. Now it's terabytes and petabytes in many applications, and the types of data are very different. It ranges from structured data to semi-structured and unstructured data.

To deal with this different environment, both in terms of the data and the way applications are built, there's a need for a new set of data management, or storage technologies, that are better suited to these applications and these types of data.

That's really where technologies like Hadoop and NoSQL databases come in. That's why they've become increasingly popular over the last few years. That doesn't mean that the Oracle database is going away or anything like that. That will still be used, and there's still applications where that's a better fit.

But, for a lot of these new types of applications that we're seeing now, these new technologies are a much better fit.

CI: I think that's where the confusion comes from, perhaps. Maybe we need more metrics around "When do I use one type of technology versus another one?"

You did a really good job of differentiating, "Look, if it's structured operational data, then the relational technologies are probably better suited to handle that kind of structure. On the other hand, if it's data that has multiple structures or has variable structures, relational databases don't know what to do with that. That's where these new technologies really shine."

Would that be correct?

TS: Absolutely. Add to that the data volume as well.

CI: And the data volume. Absolutely. We also got into an interesting, in the Chinese sense I guess, an interesting discussion on data modeling versus data schema. For me, it was a tremendous insight into the NoSQL world to understand what you meant when you said data schema or something that is schema-less, or the fact that you don't need a data model.

We all took a step back and went, "What do you mean by that?" It was an interesting discussion. It was very eye-opening. It also settled an awful lot of the confusion when someone says it's a schema-less database, or you don't need to do a data model. We're like, "How do you understand the data?" Your comments on that.

TS: There's a spectrum of applications and situations that companies have and different solutions or approaches have their pros and cons. On one hand, you could do a lot of planning and pre-organization of data and make everything very, very organized, but you'd take a big hit in terms of agility.

When you would want to import a new dataset or look at some new data, it would probably take two months for that to be prepared so you could actually analyze. On the other end of the spectrum, I just got some new raw data coming from some feed and I want to analyze the field that the developer sitting next to me just added to that log file.

In between, there's a big spectrum. We really need solutions that can address that full spectrum of use cases. That's really what it's all about.

SW: I would add that it's not about whether you have schema or not. It's when you apply that schema. Do you apply it...

CI: Yes. That was the part that became very clear.

SW: ...upfront when you put the data into a system or would you rather get the data in first, do some lightweight data discovery or exploration and then once you've defined what's of value within that large dataset, then normalize it, put it into a data modeler or schema and make it available to run your business on.

CI: The biggest 'aha' moment for me was that when you say "data model," you're talking about a schema. When I say data model, I'm talking about a logical representation of the relationships of the data to itself.

That was where it was like, "Oh my God. You're speaking Chinese and I'm speaking English… No wonder we're not communicating correctly!"

That was an interesting moment for me. It was also one where I got pushed back on my heels and I went, "Oh, now I get it. I see where you're coming from." Maybe we need to popularize that translation problem.

TS: I think also it will take time in the industry for people to get more comfortable with these new, more agile approaches.

CI: They are different.

TS: People were very uncomfortable when agile development came out and… "Where's that spec that describes what we're going to build for the next year?" That was very uncomfortable.

When NoSQL databases came out and said, "Well, you don't have to have a schema before you start using this database." There were a lot of question marks around that. Now some of these databases are among the five most commonly used databases in the world. Obviously, people are getting value.

Binks used j2ee and these kind of enterprise-y frameworks for developing applications, but at the same time, you're seeing more and more people choose frameworks that allow faster development of applications.

Technologies like Node.JS or Python and Django or Ruby on Rails, these things, they have less safety, less kind of protection, but you get other advantages in exchange for that, which are the agility and that ability to move faster.

CI: What's difficult is that we're in a period of incredible innovation of technology and keeping up with, "Do I use X, Y, Z or A, B, C technology?" I think that's where a lot of people are getting totally confused. It is an extraordinarily innovative period, which I find fascinating, but it's also pretty frightening.

Let me move to the next part, because I do want to talk about Apache Drill. You introduced that to us. My question is what is it and what kinds of features does it bring to the table?

TS: Apache Drill is scale out SQL execution engine. What it does is you install it on one or more commodity servers. You could even actually install it on your laptop.

What it allows you to do is run SQL queries on this massively scalable cluster and get results really fast, at interactive speeds. You can use the technology of the BI tool, like Tableau or MicroStrategy and just start exploring that data.

It fits that data exploration use case and makes it very much a self-service, so the end user, whether that's a data analyst or a business analyst, they can look directly at the raw data. They can start exploring it, find some insights and really find out what data they want to then centrally manage and tightly manage in the traditional sense.

CI: What was interesting to me is that yes, it will use the schema-free environment, but that's not mandatory. It will also take a schema, right? There's no barrier there.

TS: Absolutely. Actually Drill, the execution engine is built to not require schemas, but it will actually connect to different data sources, so if you have data that has schemas defined, for example, Hive tables or just data that's registered in Hive metastore, or you have files that have embedded schemas in them, or semi-relational databases, Drill can query all that data. It can actually do joins across different types of data.

In the demo that we saw...

CI: Relational as well as non-relational?

TS: Absolutely. You could have data that's relational and then you join that with log files that are JSON log files that may change because the developer an hour ago decided to add some new fields to that log file.

CI: It's not out yet, right?

SW: Actually, Drill is available. It's an open source project as part of the Apache software foundation. Its current version is 0.7 and that's available. There's a monthly release of Drill, so every month, there's a new version.

In a few weeks, we'll have 0.8 and the first half of this year, we'll have the GA, what we consider the GA release of Drill, which will be 1.0.

CI: Version one. When is it coming out?

TS: It'll some up in Q2 of this year, but we are already seeing companies take advantage of the technology in production use cases. It depends on what you're trying to accomplish and whether Drill has the functions that you already need. It's already very fast, very scalable, supports almost the entire set of SQL, various capabilities related to JSON data, etc.

There's already a lot there that can be used today.

CI: Well done. Unfortunately, we're out of time, so that's it for this edition of the BBBT podcast. Again, I'm Claudia Imhoff and it's been such a pleasure, really has, to speak with both Steven Wooledge and Tomer Shiran of MapR today.

I thought it was a wonderful session. Again, I encourage people to see the full video, because that's certainly far more interesting than just me yakking at you.

Thank you both for coming.

TS: Thank you for having us.

SW: My pleasure.

TS: It was a pleasure.

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!