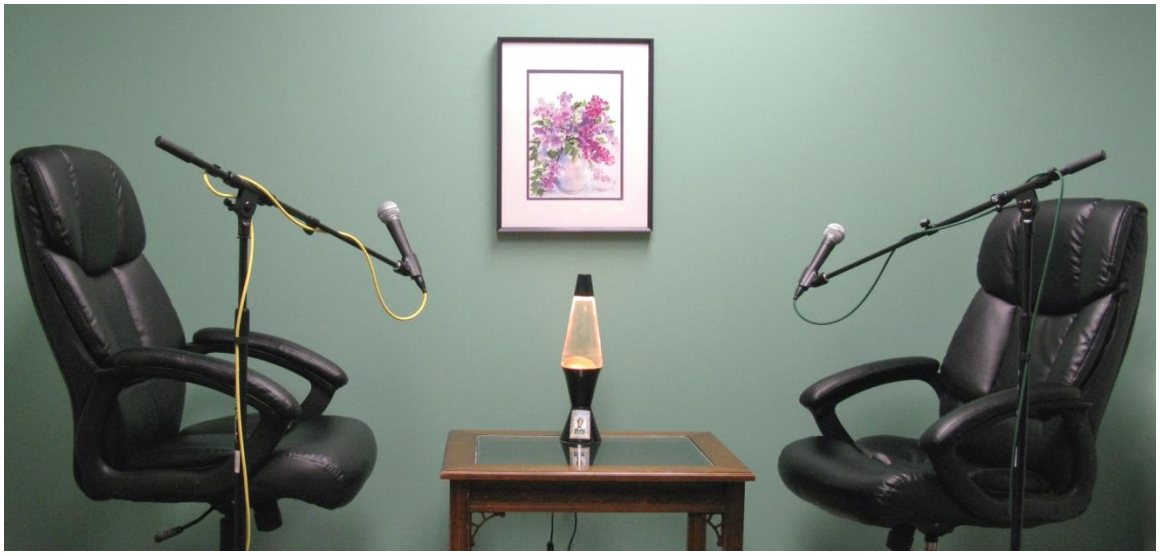
BBBT Podcast Transcript



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:	**Datameer**
Date recorded:	**August 15, 2014**
Host:	**Claudia Imhoff**, President, BBBT
Guest(s):	**Stefan Groschupf**, CEO
Run time:	**00:19:47**
Audio link:	Podcast
Transcript:	[See next page]

Copyright © 2013 BBBT - All Rights Reserved

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 is Stefan Groschupf. Stefan is the CEO for Datameer. So, welcome!

Stefan Groschupf: Hello.

CI: It's nice to have you back again. Let's start with a little bit of an overview of DataMeer. It's a relatively new company. So, why don't you talk about it just a little bit here.

SG: We are four and a half years old. The company was founded in 2009. The reason we founded the company is, we, basically, since the early 2000's, worked on Hadoop.

I joined the Nutch project very early with Doug Cutting and Andresh and built it – and a whole bunch of stuff on top of Nutch, with my colleagues together, and contributed heavily to the project.

More and more companies then came to us and said, "Hey, can you guys implement this thing on Nutch and then later on Hadoop?" We did work for Apple, for AT&T, Horizon. We predicted who's the next music star for EMI Music, based on all sorts of social signal like MySpace.

In 2009, we just did it so many times we said, "OK, there's a pattern here. We want to build this one more time, but we'll make it a product." We're very fortunate to hit a nerve there.

CI: You did, and certainly Hadoop is very popular today. There are many companies that are building the basic platform. You've got Hortonworks and Cloudera and so forth and so on. Where does Datameer, then, fit in to this overall architecture?

Copyright © 2013 BBBT - All Rights Reserved

SG: Datameer is the first product that was natively built on top of Hadoop. We bet our company that Hadoop will be big, and it turned out to be big. We really built a next-generation data analytics product that is not just connecting to Hadoop, as many, many do that have a "big data" story, but we really run everything we do, the data integration, the analytics, the visualization, on top of Hadoop. We compile into native Map-Reduce. We don't use any intermediate layer like Hive or something.

That allows us to really change some of the processes and make some really great features that significantly accelerate time to inside.

CI: Let's talk about that a little bit because there are a lot of companies -- you're right, Hadoop is very popular -- and there are a lot of companies that claim that they do this and they do that and so forth and so on. What do you see as your key differentiators?

SG: We have a fully integrated product. As I said, integration, analytics, visualization is in one product. With that comes a fully integrated security, monitoring.

But also, we do a whole bunch of very early analytics as we integrate the data that then allows us to give users data visualization, data profiling, and all that good stuff that comes. There's a whole bunch of companies that point to Hadoop and then load data out of Hadoop.

The Hadoop story seems to be, I can push data to Hadoop or I can pull it out. Then I put it into my RDBMS system and run certain things there or I maybe connect to an SQL adaptor that sits on top of Hadoop. I'm happy to talk about SQL and why it's not a good approach there.

CI: We'll get to that in a minute.

SG: The differentiation here is really that everything we do runs as a Map-Reduce job and with that we have the scale, we have the multi-source, we can work with structured, semi-structured, and unstructured data.

We are not really bound to the traditional BI product. This is more built for a single source, one DB2 or one Oracle, and you ETL the data in and that needs to be all pre-structured.

Copyright © 2013 BBBT - All Rights Reserved

It's a little bit more agile and flexible. Everything is a view now in our product. That's the biggest advantage. You don't need to predefine the structure. You still define your structure. I'm happy to talk more about the detail.

Think of it as everything is a view, like a table view in your database. That gives you agility. You can very quickly change things.

Traditionally we didn't like to use Fuse because we had limited storage and compute. What Hadoop brings to the table is unlimited storage on compute, or at least to a price that is so cheap that it does not make sense to pre-optimize data any more, but use that view concept.

What we see with customers there is that within four to eight weeks, they are really getting a tremendous amount of insight, where traditionally they had to model the data, ETL the data, build out the data warehouse environment, and then start analyzing data. Maybe you get insights or not.

This is way more agile with pulling data in, creating a schema-on-read, a view on the data in our spreadsheet user interface and getting insights and moving from there.

CI: One of the problems, though, with the view or the schema-on-read is that everyone and their cousin can create their own view of the data, which gets into a little bit of a problem. At least, in terms of two people looking at the same data but coming up with very different results.

Now we're back to this conflicting "my numbers don't match your numbers." What do you say to people when they register that? How do you mitigate that kind of a situation?

SG: It's a wonderful thing that everybody and their cousin, and the cousin of their cousin can do its own view on the data, especially if the cousin is a business user and not as technical. Again, historically, that was my life because I worked on [the] IT side.

The business user pulled a ticket and said, "I want to have a schema change," and I said, "Yeah, come back in six months," while we can't change the schema of our data warehouse. It's in production.

Copyright © 2013 BBBT - All Rights Reserved

Having that very easy way to create your own view on the data, and, by the way, data cleaning and extraction/transformation as part of that view, so that's the number one accelerator, that you move to E and the T of ETL over to that more BI business user side.

But to come back to your question, on the one hand side it's great everybody like marketing has its own view, CS has its own view, and then it's about data governance. It's not necessarily bad thing that everybody can create their own view. The question is, would you like that or not?

We have a lot of customers that very clearly use our data governance functionality to limit that. It's very clearly defined what is clean data -- and it's that one group that works on that -- and then it's very clearly defined what's a user? What's a customer for us? What attributes does a customer have? How is that customer business object relate to a purchase business object, and how the attributes [relate] there?

You can lock this down with our data governance, so it's easy. Everybody can create his own view, but that doesn't mean that everybody is allowed to create his own view on the data, and this is how you really have to deal with it.

But we have, by the way, situations where people created different views in the same data and they identified problems with an already existing view on the data. So, just having the ability to look differently at your data could be a good thing. But in the end, again, you can't have "wild west" and a "big bang". So, it's all about data governance.

CI: Well, I think what you touched on is a really important point. It used to be that BI people were touting the single version of the truth, which I thought was a terrible thing to say. There is no single version of the truth.

There is a single version of the data, but everybody has got their truth, and what you're saying is, "Look, let them create these views because that is their version of the truth."

They can always get back to the underlying data and what did you do to it because you've got the lineage and all of the tracking of the view itself. But I like your stance of "you want a different view, go for it."

Copyright © 2013 BBBT - All Rights Reserved

SG: I think what we see is that it's not so much we have different views on let's say, with the custom object. It's not so much, "I want to have a different view on the customer." It's more, "Oh, I want to have an enrichment of the customer."

You have kinkd of a, "OK, this is for my big bang." This is what a customer needs, and marketing would like to add five different things here, where sales want to add ten different things there. Traditionally, what I observed in my career in the space, was you got together in never ending meetings, you had a ginormous amount of Visio diagrams that tried to show data schemas, [and] you had a whole bunch of very non-technical business users that then tried to make any input into the data schema.

This was a month-long project. It ended in a 200-page Word documents with version 23 and what have you. The great thing is you can control [it]. You have the data governance, but you have the flexibility to adjust, and enrich, and faster iterate on things than you traditionally had.

CI: Yeah, and I like that. I think, again, it's the agility. It's the speed to insight and that sort of thing. We also had a rather heated discussion, I suppose, on some buzz phrases that are going around in the Hadoop world like "no ETL" and "SQL on Hadoop." Everybody is all hot to put SQL on Hadoop.

You had a different take on both of these, and I'd like to hear your thoughts on them.

SG: Yeah. First of all, to be very transparent, we use a no ETL buzzword as well. To demystify that buzzword, if you think about how we do data analytics, over-simplified of course since about 30-40 years, [on the] left we have [the] data source system, [and] we have ETL. In the middle we have our data warehouse, our schema, our database, our data model or whatever you want to call it, and maybe on the right hand side you have BI.

The reality of no ETL is there's no ETL. You always have to extract. You always have to transform data to make your analytics. But our product, for example, moves the "E" and the "T" of ETL over to the right hand side.

Copyright © 2013 BBBT - All Rights Reserved

It moves it to more self-service, the more non-technical user, and the reason we can do this is Hadoop has that incredible power so we don't need to pre-optimize the data anymore.

The reason we clean the data before we put [it] in our data warehouses, if I have hundred thousand records that I can't use, I'm wasting my licensing and my hardware. Right? When I started my career, I had companies spending millions and millions of dollars on database environments.

If you think about, moving to 2014, people spend millions and millions on people that know how to operate those and guess what, the hardware is maybe a few hundred thousand dollars.

The difference now is really that if we move the "E" and the "T" more on the business user and they'd make it a little bit more self-service, yeah you could make an argument it's not the most efficient way but guess what? It's not about efficient, it's about agility and how fast you get the insight. No ETL really means move "ETL" to the right hand side of the equation.

CI: And then the other one was SQL on Hadoop. How do you feel about that?

SG: SQL is a great tool. There's hundreds of thousands of people that know it, that learned it. They're heavily invested. What we have to understand is that there are a bunch of companies heavily invested SQL as well. A few companies are really pushing hard [for] SQL on top of Hadoop.

Ten years ago, when we started writing Hadoop, we had a choice and coach everybody to go back – it's public knowledge, its public email archives -- to go back into months of email conversations and discussions. Should we do SQL for our search engine or do we write our own thing?

Back then we said, "Well, we aggregate data in the Internet. We don't know yet what kind of analytics we want to do, so we don't want to pre-structure the data. We don't want to do SQL. We don't want to do the schema on write. Right?" What a database in SQL obviously requires. Where the idea then again was the schemaon-read approach. We store data first, and then everything has the view on the data we discussed.

Copyright © 2013 BBBT - All Rights Reserved

Another really important thing to keep in mind is that Hadoop, the underlying file system, is almost built to use the metaphor "like a tape drive." It's highly sequential optimized. You start writing. You write a whole bunch of data - terabytes, and then you stop writing.

It's not optimized for random access. You're looking for individual records, do a filter, do a group by - all that kind of functionality that makes SQL so great. SQL isn't great because you say, "Select * from something."

It's so great because you have a "where" clause, that then very quickly allows you to the draw-in selections and all of the good stuff.

But exactly that doesn't work very well with the underlying sequential optimized file system. So yeah, there are big companies that would like to get their products running on Hadoop because it becomes very, very popular, especially running schema-on-read.

Again, there's a conflict. If you want to do schema-on-read, you can't do SQL, and if you want to do SQL, you actually need more random access data storage like RDBMS and B-tree data structure. Hadoop doesn't have that. You're trying to duct tape the pig on top of the elephants here a little bit.

CI: Now there's a visual for you.

SG: I am not the biggest fan there. That said, there's fantastic SQL tools out there. If you need SQL, use what's out there. It's very mature. It's 10, 20, 30 years old. The optimization advantage to some of those great databases.

We are decades away in Hadoop world to even get there and if you think of a high of Hive, that is an SQL on top of Hadoop, it doesn't even support sub-selects. If you use a BI tool, one of the great BI tools out there, [like] Tableau, MicroStrategy, Qlik Tech pick your favorite one. They generate very complex SQL statements.

All those SQL, even though they claim all those SQL tools on top of Hadoop can't support those generated SQL statements, and again, the underlying file system is where, then, it all falls together.

Copyright © 2013 BBBT - All Rights Reserved

CI: Yeah. All right, we've got about a minute or so left. If you don't mind, why don't you very quickly just pick some used cases, give me a high-level overview of that, and then I want to get into a last question that was a little bit controversial?

Let's quickly go through some of the use cases, and why would Datameer fit well in them?

SG: Yeah, number one use case is customer analytics. Today, if you interact with customers, you interact in many, many chance - mobile, online, social, in your call center CRM, getting all this data is very complex.

You have structured, you have semi-structured, and unstructured data, and this is where, again, the schema-on-read the view approach, the agility is very powerful. Second biggest used case, operation analytics, getting more on structured data together with structured.

This is where we see value-based capacity planning in TelCos. This is where we see companies being predictive.

All those good things that we see. Multi-structured data and multi-source is the big topic here where we differentiate from more traditional single source, or structured data only, approaches.

CI: Last question, let's talk a little bit of another drilling into why Hadoop and when is Hadoop good and when is Hadoop not so good? One of the things that you said was that Hadoop should be used for forensic analytics. First of all, explain what you mean, and secondly why?

SG: Given the batch nature of Hadoop, you really just look into the mirror, you look back in the mirror. Hadoop isn't great, if you want to do auditing or real time decision-making because it's always batch.

Sure you get the batch down to maybe just a couple of minutes, or maybe, with the next generation of technologies to 30 seconds, but you always look at historic data.

CI: There's a certain latency built-in.

Copyright © 2013 BBBT - All Rights Reserved

SG: It is, and you will not get around that except [if] you can break the laws of physics. Some companies say they can. I question it. So, Hadoop is great to look at historical data, especially if it's large data. If you have small streaming data, there are other technologies that are way more suited for that.

CI: And I liked the other thing that you said, that Hadoop is really good for experimental types of things -- maybe you don't know what you want to ask until you sit down -- whereas the enterprise data warehouse is not going away. It's going to stay there for a while, but it's more of a production standardized kind analytic. In other words, I am going to run the same report or I am going to do the same analytics. I want to put that in a production environment. If I want to experiment then that's where Hadoop really shines. Right?

SG: Absolutely. Where our product is heavily used is data discovery, really meshing together all sorts of data sources - structured, semi-structured, unstructured, and trying to get those insights and then productionalize that.

Where RDBMS really shines is the traditional financial BI reporting. Every Monday at 6:00 I have exactly the same report on my desk. You can do this with Hadoop, no question, but where you get this dramatic improvement of time to insight is the more data discovery approaches.

CI: Yeah, and it's a more appropriate usage. If you're going to go down that road, use it for what it's really good at.

SG: Yeah, that's what we built it for.

CI: You have a very bright future. You are perfectly positioned, then, I think.

Unfortunately, we are out of time for this edition of the BBBT podcast. Again, I am Claudia Imhoff. It's been a great pleasure to speak with Stefan Groschupf of Datameer today. Thanks so much.

SG: Thank you so much for your time.

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

Copyright © 2013 BBBT - All Rights Reserved

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!

Copyright © 2013 BBBT - All Rights Reserved