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:	**DataSift**
Date recorded:	**August 14, 2015**
Host:	**Claudia Imhoff**, President, BBBT
Guest(s):	**Jason Rose**, Senior Marketing VP
	Ben Hagan, Principle Sales Engineer
Run time:	**00:16:27**
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 Jason Rose and Ben Hagan, Jason is the Senior Vice-President for marketing, and Ben is the principal Sales Engineer for DataSift. Welcome to you both.

Jason Rose: Thank you, Claudia.

Ben Hagan: Thank you. It's nice to be here.

CI: All right, Jason. Let me start with you. First of all, you stated that DataSift is a human data company. Not many people know what that means, so let's just start there. Why don't you define what that means, and what the sources and volumes of this human generated data is all about?

JR: We live in a digital age, where people are posting and putting their thoughts and feelings online more than at any time in history. DataSift really operates in that realm of digital human-created data. If you think about social networks and platforms, like Facebook, Tumblr, blog solutions like WordPress, which powers something like 20 percent of the world's Internet sites, news sources like LexisNexis, which has a compendium of around 250,000 different news sources. Those are the sources that DataSift deals in.

In terms of volumes, we're talking billions and billions of items every day. DataSift provides a platform for you to go from those billions and billions of items, most of which likely don't apply to your business, just to get down to those few pieces of information that you want to either have information about or take action on.

CI: I feel like Carl Sagan just stepped in with his billions and billions there. Let me stay with you on this question as well, Jason. A little bit about DataSift.

It's been around for what, four years or so? A little of the history of the company... that would be very useful, and a little bit about what its architecture looks like as well.

JR: OK, great. We're a little known company. We started off in a town outside of London called Reading, in the UK. Our founder, Nick Halstead, started off a company called Favorite, which for those of you that remember RSS feeds, it was an RSS feed curation tool.

Then we saw this new little company called Twitter. We started to realize that we could surface new stories from that Twitter data.

We started a company called TweetMeme out of that, where we realized it was really challenging for people to share information on Twitter. We invented a little thing called the Retweet button, which we embedded in about 500 million different websites. Then we had the bright idea that the value was in the data.

We made a deal with Twitter, where we traded the intellectual property around the Retweet button to Twitter, in exchange for the full fire hose. DataSift was born. Our core tenant has always been about the technologies. We've built a generalized platform that really can handle any of that human generated data that I spoke about earlier.

CI: You've gone way beyond Twitter. Just to name a few. You've named some of the... LexisNexis, and so forth, and so on. There's also a relatively large gorilla that you're now working with. We'll get to that in a moment, but who is that?

JR: We announced a partnership with Facebook, who some of you may use and may have heard of. There's about 1.5 billion people currently on Facebook, approximately half of all Internet users use that social platform. We are a very proud partner of theirs.

CI: Excellent. It seems to me, and I think you confirmed this, that one of the biggest challenges in dealing with human generated data is making sense of it all. You mentioned the billions and billions of items, tweets or whatever that Facebook postings and so forth.

Finding the handful that are actually useful means that the data does have to be curated. It has to be put into some form or format, where it can then be analyzed. Would you agree? How does DataSift go about doing that?

JR: That really is DataSift's core mission. Most companies, when you talk about unstructured data, get a bit of sour taste in their mouth. I don't think there's very many companies that are really successful at getting value out of unstructured data. This digital universe of human generated data is highly emotive and highly important, but also very noisy and very voluminous.

The core of DataSift's technology is our underlying data model. We take all of those different sources that you mentioned, and we put them into a normalized and enriched data model. For example, if there's a short link in one of the posts, we unravel that link to find the original URL.

Let's say for example "The New York Times," we bring back all the metadata around that page, so the title of the article, who the author was and any other keywords or hashtags around that, but of course, this human generated data is also very quickly moving. We do all of that. We add sentiment, topic extraction... I could go on and on.

We do all that in under 300 milliseconds per item. It's coming through in real-time, and we're providing incredible insights into that information, and providing a single way for customers to query, filter out the noise and get down to just those few pieces of information that are actually relevant to their business.

CI: You touched on that. I want to touch on it a little bit more. It's more than just taking unstructured data and making it structured, because what DataSift as I understand it does, it also adds a tremendous amount of richness... context. When I say something is bad, do I mean that it's really bad, or am I using slang and so forth, and so on? It's more than just sentiment analysis. It's the whole context that I'm using these terms in. Is that correct?

JR: That is correct. It's a number of different aspects. One is topic and entity extraction, to figure out what people are actually sharing and talking

about. There's also you mentioned sentiment, so the ability to score and look at what people are saying... positive, negative, neutral, rant, rave. Then we also do what we call intent classification, where we have a sophisticated technology called VEDO, which allows you to use machine learning, but in a very simple way.

You classify a training set. Our machine learning builds an algorithm to classify the rest of the training set at scale. It's actually smart enough that if it finds something that's a little ambiguous, it'll feed it back to you as a user, to say, "Can you classify this for me, so I can learn better and do a better job of classifying all of that information at scale."

CI: All right. Let's get into the technology a little bit, then I want to bring you into the conversation. Let's start off with what you call CSDL. It is not an acronym that most people will know. What does it mean, and how do I use it?

BH: Sure. CSDL, first of all, stands for Curated Stream Definition Language. That's definitely bit of a mouthful, but it's the heart of the DataSift platform. It's been something that we've developed over a number of years, and it's the filtering language that allows you to take the billions of items that Jason described earlier, and focus those down to the core pieces of content that of relevant for you.

It's reducing the noise. It's taking out the potential spam. It gives you the ability to be very precise around what data it is that you're interested in... what's collected. So yeah, the CSDL is really the heart of the filtering platform in that language. That's actually accessible programmatically as well. You can generate these filters on a programmatic basis to automate the creation of this.

CI: Excellent. All right. He has mentioned VEDO. You also have VEDO Intent. Let's step into those two.

BH: There's two problems that we solve here at a very high level. One is getting the right data. That's where the core CSDL functionality comes in, and it says, "Is this data relevant to my use case and my business?" The

second part of that, and this is where VEDO comes in is understanding the context of that data.

What is it that people are saying and getting understanding in meaning and deep signal? There's a number of features within VEDO to support those types of rules and classifications. Jason has already talked about how we're starting to leverage machine learning. We bought out our products called VEDO Intent, which makes that easier to bring into your use case.

It makes it easy to generate classifiers and to use that machine learning technology. We're really working hard to bring those more complex data science problems, and make them simpler and more consumable and easier for people to consume.

CI: What I liked about it was the fact that as you said, Jason, the human can get involved. Confusion can reign. “We're not quite sure what to do with this. It's ambiguous. I don't know if this is a rant or a rave.” Your technology basically says, "I don't know what to do with this human. Get back in here, and tell me what to do."

BH: Exactly. It's an iterative process, where the platform is continuously looking at the classifications you've given it and comparing those to its expectations. It will give that back to the human if it's unsure of certain areas, and it will give you readings and outputs on how successful your classifications have been.

CI: Excellent. All right. Let's talk about the last one that you covered, and that's PYLON.

BH: Yes. PYLON is a brand-new product for DataSift. This is really changing the way that people can access data. PYLON is an analytics product. Our first data source with PYLON is Facebook Topic Data for PYLON. What we are actually doing here is DataSift is embedded inside of Facebook's architecture.

What PYLON effectively does is gives us a privacy-first approach to accessing anonymized and aggregated data. It's allowing people to start

understanding what's being said at a high level, without being allowed to drill into individual users.

CI: All right, Jason. Let's go back to you. Let's finish up with you. It is a very important change to your relationship. You now have a very deep relationship with Facebook. In fact, exclusive, if I can use that word.

Tell me a little bit about what DataSift is doing with the Facebook Topic Data. As been mentioned, one of the most important parts of that is this privacy-first management control that you have. So, let's touch on that, because that's important to a lot of Facebook users.

JR: Absolutely. We are the only provider of something called Facebook Topic Data. As most people know, Facebook is a private social network. It's unlike, we mentioned Twitter earlier, Tumblr, a lot of these networks... Anything you post becomes immediately available to the entire world, so there are different privacy considerations with those networks.

With Facebook, most people set their settings to be friends or friends of friends, so you don't expect your Facebook content to be broadcasted worldwide. We've worked very closely with Facebook, to innovate this new product that Ben mentioned called PYLON, that basically allows us to take a privacy-first approach to private networks, like Facebook.

Basically, what we've done is we've installed our software on premise inside Facebook's data centers, so Facebook maintains control and ownership of topic data throughout the entire value cycle. What happens is we are allowing customers to set up a query into all the posts, shares and engagement inside the Facebook network across 55 different countries in a wide variety of languages.

What we do is collect that content, but before we collect it, we, for example never receive any user information. The data is always maintained and stored inside Facebook's infrastructure. It never leaves Facebook, so to speak. We actually do a fair bit of work to make sure that any result that's provided back is aggregated and anonymized, so no personally identifiable information is revealed.

Within that, we receive a really strong set of demographic information that allows end users to understand their audience. What are they sharing? What are they engaging in? Where are they? How old are they? These are self-declared demographics. When you sign up for Facebook, and you say, "I'm such and such. I'm this age, and I live in this location," those details are pended to the information, but then aggregated up, so that we never know individually who is talking about a particular subject, but at the aggregate level, we can understand the audience.

CI: Yeah. I found it fascinating, because the way a brand manager perhaps, would use that information, let me let you explain. How would I use that information?

JR: It's a great question. Lots of marketers will create all kinds of different campaigns. They want to know if I use subject line A or subject line B. Am I getting more uptake and interest in that particular offer? Previously inside Facebook, you could always see what was going on around your brand page.

I could always know that the DataSift page on Facebook, if you came up to it, liked it, shared content from it, I could understand that. I could see that. However, if you started off a conversation in your news feed about DataSift, I would never be able to understand what's being shared and engaged with.

Now, if you think about the broader sharing and engagement on Facebook, I can get insight into that, so I can see which URLs, which hashtags and what content is being shared and engaged in. I can begin to tailor my marketing messages, my new product that I'm about to introduce, the features in that, or make sure I'm using the right language to connect with my audience.

CI: Your technology is embedded mostly. You gave us some models of who uses your technology. You got about 30 seconds. Tell me the models of who uses your technology.

JR: Our go-to-market, if you look at us, we are an API, basically. Our target market is developers, for the most part, and what I would call lead

adopters. As you probably heard through this conversation, we've been talking about some pretty heady, big data sources.

We really are appealing to what I would call early adopters, the Sysomos the Nubes, Synthesios of the world that are building products on top of our platform, and then going to consumers, versus us going direct to an enterprise or a consumer.

CI: It's an interesting world you're in. Unfortunately, we're out of time right now. Thanks so much. That's it for this edition of the BBBT Podcast. Again, I'm Claudia Imhoff. It's been a very great pleasure to speak with Jason Ross and Ben Hagan of DataSift today. Thanks again for speaking with me.

BH: Thank you.

JR: 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!