



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:	Snowflake
Date recorded:	October 16, 2015
Host:	Claudia Imhoff , President, BBBT
Guest(s):	Bob Muglia , CEO Jon Bock , VP of Product and Marketing
Run time:	00:16:54
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 so pleased to introduce my guests today. They are Bob Muglia, and Jon Bock. Bob is the CEO, and Jon is the Vice President of Product and Marketing, for Snowflake. Welcome to you both.

Bob Muglia: Glad to be here, Claudia.

Jon Bock: Thank you.

CI: All right, Bob. Let me start with you. Snowflake's a relatively new company. It's about three years old, or thereabouts. If you don't mind, give me an overview of the company.

BM: Sure, I'm glad to.

Three years ago, our initial two founders, Benoit and Thierry—who are architects at Oracle—they were maintaining the Oracle database, and building Exadata features.

They realized, in looking at what they were doing, that there was a different way that was possible to actually build a data warehouse using Cloud technology to solve problems for customers in ways that just wasn't possible before the Cloud came around.

They decided to form a new company. They were joined a few months later by Marcin, who is our third founder, and was the CTO of VectorWise. His PhD paper was in database vectorization, so he's an expert wiz at making data warehouses run extremely fast.

They built this data warehouse from scratch, a fully relational SQL database, not built on any existing code base. It's not built on PostgreSQL



or MySQL or Hadoop. It's a full, SQL relational database that they wrote that's designed and built to run in the Cloud environment and solve the problems that our customers are seeing today.

Today's real world problem is solved in a very different and more effective way.

CI: Jon, let me bring you in to the conversation. What are some of the reasons for customers to choose Snowflake?

JB: Absolutely. Some of the most common things that customers are seeing when they come and talk to us at Snowflake come down to... one is they have this really complex infrastructure they're trying to deal with for processing data.

They have a bunch of different data marts, maybe a Legacy DW... They might have some different NoSQL systems, and they're trying to figure out how to simplify that.

Another challenge they have is just that this costs a lot. We have a lot of customers who come to us because they're facing multi-million dollar costs to actually expand their capacity for data processing.

Then, in addition to that, they're dealing with both data changes and changes in what users are trying to do. They're dealing with semi-structured data, with structured data, trying to bring that together.

They're trying to make that available to analysts faster, and trying to deal with the fact that analysts are struggling with, "It's running slowly for me today because someone else is busy on the system." Or, "I can't get the data fast enough to really make use of it in the way that I want to make use of it."

CI: Bob, back to you. Let's turn to your product. Why don't you describe the characteristics of your product, if you don't mind.

BM: Sure. As Jon was saying, customers are seeing new challenges in data warehousing needs. In particular, one of the things they're seeing is new



data types, like semi-structured data, JSON, in addition to traditional transactional data.

There's really two kinds of data people are working with—transactional, as well as machine generated data. What our founders realized is that by using Cloud technology in a different way, we could build a solution that spanned both of those and did so much more effectively.

For as many years as I can remember, people have been saying that you need to bring—in terms of data processing—the compute to where the data is. This was true before the Cloud came about.

Really the combination of, essentially, unlimited BLOB storage in things like S3 in the Cloud, virtualized computing—instantiating compute on demand—as well as really fast Ethernet networks, have allowed us to build a brand new product that fully separates the storage from the data, as well as from the brains.

In some ways, we have three layers to the architecture. There's the actual data storage, which is done in BLOB storage together with a metadata store. There's the compute processing, the actual processing of the data that happens inside a virtual warehouse. Then, we have a services layer that sits on top of that. It bridges the two and does all of the query processing, and allows us to process the data extremely effectively.

That new architecture is designed to work with both structured data that's out of transactional systems, as well as semi-structured data—data in formats like JSON and Avro.

Unlike anything else on the planet, we can take and put a full relational engine across all of those, and allow you to do standardized SQL queries that work against this broad set of data. It really results in some incredibly different characteristics, and that all is based on this architectural separation of storage and compute.

CI: If you don't mind, give me a little more detail of how you store the data.

BM: Sure. We're really smart. First of all, what we do, is we use columnar compression in the way we store all the data.



When you load data in to Snowflake, we immediately columnar compress the data and then we collect statistics—metadata about that. We do that for both structured data as well as semi-structured data.

One of the very interesting things about Snowflake is we actually don't require you to define a schema for semi-structured data. You can just load JSON data directly into something we call a variant data type. It's a special data type, that we define, that allows you to work with this semi-structured data and then apply relational processing—standardized SQL queries—on top of it.

While we don't require you to declare any structure to that data, we actually infer the structure as it's being readied. We see attributes that are repetitive, we columnarize those attributes to store them efficiently, we collect statistics. Then when you run queries against that, it's incredibly fast. We're able to prune out the data you don't need.

One of the interesting things about the Snowflake architecture is, if you look above the layer of storage—that combination of BLOB and metadata storage—we have separated out and built this virtualized compute architecture.

With Snowflake, those two are separate. Storage and compute are separate. You can instantiate as many virtualized warehouses—compute clusters, that can be created on demand—to work against the same data at the same time.

Unlike a traditional architecture, like a Shared-Nothing architecture where the data is tightly coupled to the nodes, in Snowflake it's completely separated. You can have multiple clusters, multiple virtual warehouses, working simultaneously against the same data.

We've maintained transactional coherency across all of that. Writers never block readers and you always see a consistent view of the data.

What this enables is some very unique things. You can have a virtual warehouse loading data 24/7, and that loading process has no impact on a separate virtual warehouse that users are using to run queries. You can



have multiple warehouses running queries at the same time, again against the same data.

That's a unique architecture that no one else can deliver.

CI: Just to be clear, you are only Cloud based?

BM: Right. We're entirely based in the Cloud. We're actually a "software as a service" offering. We're like Salesforce, in that sense. You acquire an account with Snowflake... you load the data, you run the queries, we do all the rest.

One of the things about Snowflake is we get rid of all of the manual tuning that's required. It's been designed to be run in a way which does not require a significant amount of database administration, in fact, almost no administration at all. And, we run as a Cloud service.

CI: Let's talk a little bit about the Cloud service. One of the concerns that many potential customers have of Cloud-based services is security. I think you have a very robust architecture behind your security. If you don't mind, talk about that a little bit.

BM: Customers care about their data a lot. They care a lot about making sure that their data is secured, and they want to have a lot of assurance of that if they're going to put their data in the Cloud and work with a vendor like Snowflake.

We've focused, from the very beginning, on ensuring that the system is highly secure. Probably the most important thing, here, is all data is always encrypted in Snowflake, all the time. We maintain the key structure for that automatically for our customers. This is a complicated and difficult thing to do, but it provides a great deal of increased assurance in security.

We also support a wide variety of features like complete role-based authentication, so users have different views in to the database... and the roles they're in. This is actually looks pretty standard, relative to what someone would see in a traditional data warehouse.



We've extended that to include the virtual warehouses themselves, and different users have access to different virtual warehouses.

We support a wide variety of other features like two-factor authentication, and federated authentication. We're in the process of HIPAA certification. We take this incredibly seriously.

We often have situations where enterprise customers say to us, "I'm worried about putting data in the Cloud. Tell me about your security." When we spend half hour, an hour, with them they come out feeling very much assured.

CI: Your security is probably much better than their own, right?

BM: We've taken it very seriously.

One of the things that's very helpful, I'll say the following, is we have one service that we need to secure. That's our data warehousing service. The enterprise customers we deal with often have hundreds, or thousands, of applications that they need to secure. So ,in some senses, their problem is more complicated.

CI: Indeed it is. All right, Jon, back to you. There are competitors, you have a number of them. I'm just curious, what do you see as the major differentiators from these other companies offering similar types of things?

JB: Fundamentally, it starts with the architecture. There are definitely a lot of offerings today that are available in the Cloud, that took a traditional database that was designed for on-premises solutions, and have put it into the Cloud with maybe a little bit of Cloud infrastructure glue.

We actually stepped back and said, "Let's re-design the data warehouse specifically for the Cloud." The architecture that we designed gives us a number of capabilities that you just wouldn't find if you took an existing database and put it into the cloud.

Bob already talked a little bit about the fact that you have this separation of compute and storage. I could store massive amounts of data, but only



have to have the amount of compute that I need, and I can have multiple independent clusters.

Instead of having different users fight with other users over who gets to access the data when, and who's going to slow down who's job, we have a technology solution that takes that question off the table and that concern off the table. That's certainly another part of key differentiation.

In addition, that semi-structured data capability that Bob was talking about... the ability to take what used to be data that you had to pre-process before you could put it in a SQL database, get that in as-is, query it using SQL, and get great performance—that's a clear differentiator that is not something you find in a stock relational database.

Certainly the performance you get there is better than what you'd find even in a NoSQL database for analytic querying.

All of those are fundamentally related to the architecture. On top of that is this whole notion of delivering it as a service.

When you take an existing database and you put it in the cloud, there's still a lot of stuff you have to manage. You might have to manage indexes, you might have to manage partition keys, you might have to manage copies of data because you're trying to create multiple clusters and have them have the same data.

All of that—security, availability, disaster recovery—that's all part of the service. It's just built in. You get that just by nature of using Snowflake without having to do extra work on your part.

BM: One thing to add to what Jon said is also the ability to work with both structured and semi-structured data, and to be able to bring those two things together.

Customers who have JSON data, and they want to process it, very typically what they do today is they use Hadoop to do some sort of pre-processing on that data. They quite frequently, ultimately, put it in a relational data warehouse where they have data analysts, and business analysts, working against that data.



With our solution, we can take that entire process and completely streamline it because we can do native-relational processing and SQL queries against semi-structured JSON data.

A user can be sitting there working with a BI tool—a tool like Looker, or Tableau, MicroStrategy, whatever—and connecting to and working with data that is a combination of transactional business data and data that is coming out of their machine generated systems.

No other product on the market can really deliver that. That's another major differentiator, and it's been one that has been of a great relief to customers that were struggling with alternative solutions.

CI: You have a terrific analogy for how a user would set up their data warehouse, size-wise. If you don't mind, just explain a little bit about the sizing of the data warehouses in your analogy to a certain piece of clothing.

BM: When we looked at the sizing of how we structured our sizing structure, we have these virtual warehouses and you can create them at any time on demand, and you can change the size of them on-demand, literally while queries are running.

If you're running queries in a virtual warehouse and you want to make it larger, you can change the size. The next query that runs will actually take advantage of the larger size.

We want to do something that's a little bit more abstract, so we use T-shirt sizing. It's small, medium, large, extra-large. Essentially each one of those size increments doubles the size of the virtual warehouse. In other words, twice as many nodes are in the cluster as the one before it.

CI: Pretty linear performance?

BM: It's linear. The performance is virtually linear. Right now we go up to two extra-large, which is quite a large virtual warehouse. We're shortly going to introduce three extra-large, which is really an incredible amount of computing capacity.



Although we can go bigger, we've tested larger in the lab, but we just don't see customers that have a need for more processing capacity at this point.

CI: Some day you may see that. Who knows?

BM: We'll see. It will happen.

CI: Jon, let me end up with you and just an overview of a few of your customers, if you don't mind.

JB: Absolutely. Building on that point that Bob was making about the fact that being able to process semi-structured data and structured data in one place is really important for a lot of customers, we have a gaming customer, Kixeye.

They do massive online games. They have a huge amount of this event data coming from their games in real time that actually is semi-structured data—it happens to be JSON, in their case. For the programmers who are writing the games, that makes it much easier for them to generate that data.

In their case, they were struggling because they were trying to figure out, "How do I take that data and get it in to a data pipeline that allows me to then do SQL analysis on that, in combination with some of my relational data?"

At first they were looking at a solution that included a Hadoop cluster and some additional databases, and some transformations, and it just wasn't the right fit for them. It took a lot of work to keep that up to date. It wasn't fast enough to get the data to users who wanted it, in the time that they needed it.

They came in and looked at Snowflake as a way to go about providing a better solution for this data pipeline.

They've been able to take what used to be multiple systems, multiple steps, several hours to get data, from when it was generated, to an actual user...they've been able to cut that down to a few minutes by putting



Snowflake in there because we could natively handle the semi-structured data and combine it with their structured relational data as well.

That's an example that's a little bit more of the types of companies who have this new data. There's also use-cases that make a lot of sense for customers who have a more traditional environment.

We've been talking with a customer who's a large telco provider. They have an environment where they have a challenge in that they have a huge amount of data coming in, they need to get that to analysts, and today those analysts are complaining because it's hard for them to get the queries that they need to run done in the time they need.

A traditional solution would be to add a bunch of different data marts to the environment. That gets more and more complex to manage as you get more data and more users.

In their case, they can take Snowflake. Because of our architecture, they can actually have different virtual warehouses for all these different users and groups. They can effectively meet the SLAs that those users want to see for performance and concurrency that they wouldn't have been able to meet in the past, short of a very complex, costly solution.

CI: Those are beautiful examples. We could talk forever, I think, but unfortunately we're out of time now.

That's it for this edition of the BBBT podcast. Again, I'm Claudia Imhoff and it's been such a pleasure to speak with Bob Muglia and Jon Bock of Snowflake, today. Thank you both!

BM: Thanks a lot! Appreciate it.

JB: Thank you.

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!