NETWORKWORLD
FROM IDG

TD Ameritrade's big data push 1 yr. later: Benefits coming from all corners

Data quality is improving, personalization capabilities are emerging, and the pace of innovation is on the rise

By John Dix
Editor in Chief, Network World | JAN 25, 2016 1:51 PM PT

Credit: Ameritrade

Network World Editor in Chief John Dix first spoke to Derek Strauss a year ago when he was about three years into his new new role as TD Ameritra*de's first Chief Data* Officer. He had built a new group, the Enterprise Data and Analytics Group, and just finished 18 months of work to stand up nine new platforms, including a Hadoop data store and a metadata repository. Dix recently visited Strauss to see how this massive undertaking is working out.

Derek Strauss, Chief Data Officer, TD Ameritrade

Where do we start for an update on what you've achieved since we last spoke? I've got a long list of things we've been tracking in terms of value, so I can hit some of the high spots, and then it might be good to step back and look at some of the other things we're gearing up for that are only possible because of the foundation we've laid. We're going to be embarking on a pretty aggressive timeline for these new initiatives, and I feel good about being aggressive because the foundation is in place. You mentioned the Hadoop effort so why don't we start there. The drive with Hadoop is around personalization so our clients feel like we know them and we can provide useful insights and education without it feeling creepy. The focus is to be like Amazon's suggestions, where you go, "Wow, I like what they're suggesting, that's really useful."

We're calling the Hadoop environment the data marshalling yard. Why? Because that's what is typically upstream from a warehouse. Think about raw materials being brought together to be manufactured into something. They will often be transported by rail and come into a marshalling yard where they'll be sorted for delivery to various factories and warehouses downstream, and you perform analytics on the raw material as it stands. So it seemed like a natural analogy to call it a data marshalling yard. What have we done with that? A couple of key things. We have mainly focused on pulling in chat information and emails, a lot of textual stuff, to try and understand client behavior and so we can optimize the client experience in terms of scenarios. We're also looking at what our clients are talking about and reading. When they phone us, what do they want to talk about? Putting all of that together with their activity on our site, we figure out this client is really interested in certain types of asset classes and we can then look to see if there any reports by third parties, by government, by whoever, and say, "It seems like this is an area you're interested in. Are you aware these resources have just

been published and here's a link to them." All of that is around personalization.
So we're realizing analytics benefits, but there are also benefits around data and data management.
Let's take a simple example of a codes table. A code could be anything, but let's look at country codes. South Africa is ZA. USA is United States of America. When it comes to programmers writing programs, if there isn't one country code table everyone can refer to as the authoritative table, everyone hard codes the table into their program. But any large organization has hundreds of systems, so you've probably got 100 country code tables hanging around, or worse, one for every program.

Master data management is all about trying to solve that. Country code is just one simple example, but when we started looking at this it was amazing how many times people have created redundant tables, and that can lead to all sorts of regulatory and compliance problems and a lot of inaccuracies.
Take me, for example. I was born in Rhodesia. Rhodesia doesn't exist anymore, but if you're looking for Derek's birthplace, are you going to know Rhodesia is now Zimbabwe? Keeping that memory of geographical stuff centralized is something every organization needs and no one really has.
We implemented a master data management capability and the first thing we tackled was country codes. Now our application development teams know they can go to one authoritative source to find it. They're not continuing to perpetuate the redundancy and the inaccuracies in the data, plus if something changes, they don't have to remember to update their program because someone in the business now owns and is responsible for updating that data.
Those kinds of efficiencies are huge and very often get overlooked. When you think of the Chief Data Officer role, people just think about the sizzle of the analytics side, but there's a very real efficiency side on the data set which is a big plus for any organization.
Once you have this master data management capability, I presume you go around looking for duplication of effort and multiple versions of the truth?
Right. And when you find it you need to find someone to own it. That's the data governance side of things. You find an owner and that owner points to the data steward who is normally someone who is already doing work trying to fix the problem, and you say, "Here's a tool where you can analyze all the different values you've got today, harmonize them, create one source of the truth and you own that and you make sure that is up to date and everyone else starts using that." That makes a big difference. But there are literally hundreds and hundreds of instances where this would apply and it's a question of working with the business groups who are constantly tripping over these things, prioritizing them, and just picking them off one at a time and working through it.
The big elephant in the room is the client, because we, like many financial organizations, have grown up being account-centric. So John, let's open an account for you. Oh, and you'd like to try something else? Well, let's open another account for you, and another, and another. Every time we open an account for you we redundantly create information about you in that account record. We don't have one central record about you.
Behind the scenes, for financial firms to be able to deal with you as a client and

understand your total business with us and treat you accordingly, we've got a thousand gnomes running around all night trying to bring all this information together.
I'm exaggerating for effect, of course, but it's a big thing because it's like open heart surgery for the organization and you've got to really know that you're going to be successful and you've got to plan the creation of a client master very carefully. We now have an opportunity to address that head-on because we've put a lot of the building blocks in place. I'll come back to that one. That was just sowing the seed. Master data management is a key benefit and it's all about efficiency.
Data quality improvement is another key benefit. The Patriot Act stipulated a bunch of things about anti-money laundering, and there are about five major attributes of client that are critical and have to be in good order. One of them is date of birth.
How could there be any fluctuation around that?
Any company that has grown through acquisition has had to make some decisions where expediency won out over guarantees for the highest quality of data. For example, if we had acquired a book of business with a couple thousand clients and their records related to date of birth were incomplete, we might have decided to bring them in with today's date being the date of birth and the idea that we would go back and fix it over time. The expedient thing was to get the conversion done. Other times the programs capturing the data in the companies we acquired didn't have the right sort of edits so you had people with birth dates in the 1800s instead of the 1900 or birthdates in the future. Just crazy stuff.
We saw all those things and thought, "Okay, this is going to be interesting. We're going to have to do some real work analyzing these and figuring out the root causes and figuring out the best way of remediation."
In the past we didn't know the extent of the problem. We stumbled on it occasionally and have had problems running various types of reports, and we've had to rush back and try to figure out what was going on. Now we know what's going on. Now we know where the problems are. Now we're actually going back and working to fix it, which is huge. That's all the authorities want from any organization they audit. They know it's not perfect. It's what you're doing about it and do you understand the risk.
And all of these things, of course, have spinoff advantages to the analytics group because they're starting to work with data that is in better shape, and of course if you're working off data that's got high integrity your decisions are going to be stronger and it's going to be easier.
Are you bringing all the data into one place to improve the quality, or trying to improve it where it sits?
We're trying to fix it where it is, at the actual source. But that's a good point because, as we start thinking about creating a client master, ideally in the fullness of time we'll have just one place where that data is and it will be good data. But because we've started fixing it at the source now, when we do create that client master we're going to be creating it with good data as opposed to data that we have to go fix.
But its complicated. If there are seven different sources for this particular thing, say, for date of birth, which of those would we consider to be the authoritative source? If we really wanted to save ourselves the trouble of trying to fix all seven of them, which one would we fix now? We're trying to do that thinking as well.
In some cases it's not possible to do that; we've got to go out to all seven because of

the way our systems are set up. But in other cases it's possible to just go after one now. Again, this blocking and tackling around data wrangling is not the sexy stuff, it's not the sizzle, but it's critical to getting it right for the organization.

Has all of this effort required you to bring in some new types of specialists?

We're not going to employ 100 data scientists. It's just not going to happen in a company our size. It's much better to try and think of a way to crowd source our data science skills.

So working with some universities we set up a collaborative data science platform using an Amazon Cloud. We moved a bunch of our data up there, signed NDA agreements with about 12 universities and said, "You guys need real data so your masters and your doctorate students can roll their sleeves up and play with data, and we need crowd sourcing of ideas. This is a marriage. We can both give and get something from this." We had a formal launch of the platform in June and we've had really good interaction between our analysts and the university guys. The universities have come back with phenomenal ideas and insights that we're still developing. Over time it gives us access to some of the best and brightest students, some of which may want to come join us. This has been very successful and we continue to push.

Coming back to the client master, where do you stand in creating that?

We created a client profile from a lot of the data we've been collecting, which is a consolidated view of key client attributes. We've never had a client record as such and this is a start, but this is not the master yet. This is tactical, but we're already starting to use that to effectively target specific clients because we now have a view of what their interests are. In fact, this is part of the bigger personalization initiative.

Within personalization there may be 20 different topics. One of them is onboarding. When we onboard our clients we're creating 30 attributes related to that client and right now we're holding it in an Oracle database, but in time we're going to set up our client master and move this into the client master domain.

So, you will still have multiple versions but now synchronized?

It will take some before it's one and only one that everyone is using directly. Usually what happens is you first create what's known as a registry, which is a central index which creates the joins between all these different instances where your client records are held. You'll start using that as a point people can refer to, and it grows over time and you're creating more and more authoritative data in it. It refines over time and ultimately it becomes the golden source, the golden record everyone uses. It's a journey. It takes a couple of years to achieve that, but the registry, the client index, is something you can stand up much faster.

So there are interim steps toward that Holy Grail.

Yes. There's certain data our business folks have wanted to get their hands on forever, and for one reason or another it's just been too hard to get hold of. We've now implemented this virtual capability where we don't have to move the data. We can actually create a view of the data across many different sources and that has helped people get an understanding of the data without having to write new programs.

In the past, someone in analytics would say, "In order to do this I think I need this kind of data and I think it's sitting in those systems." Then they'd go to the data warehouse team and say, "I need that data to be extracted, transformed and loaded into the enterprise data warehouse."

That would take three to six months to do. And only when the data was in the warehouse could they determine, “Yeah, this is good stuff, this is what I wanted,” or, “This is not what I thought it was.”

So now, before anyone makes a request of the data warehouse, they must first have created a virtual view which takes days as opposed to months to do and they must analyze that data as it stands and determine whether that is actually the data they want.

If it does prove to be data you want, then you have to go back and submit a request in the traditional manner?

You’ve got a choice at that point. The virtualization software has the ability to cache the data so you can add it to an existing data warehouse record, or you can say, “I’m absolutely convinced this is something we want to have perfected through the extract, transform and load process and we’re going to do that through the classical route.” So you can either append or build it into the plumbing.

That has saved a lot of time because you know how it goes. There are always myriad requests to the data warehouse, and those folks get swamped and then you have to prioritize the requests and some get pushed to the bottom. Now we have a tool that lets people mine for nuggets, lets them prove the nuggets are there.

Is it self-service?

Yeah. It’s pretty cool. There’s a bit of hand-holding, but over time it’s going to be more and more self-service as their skills get stronger.

Have all these new capabilities led to some creative new thinking?

Very much so. The analytics teams are in each of the lines of business, and now they’re getting their hands on the data with these improved tools and starting to see these pictures emerge, so they’re coming forward with a lot of the ideas. So it has absolutely increased the rate of innovation.

We’ve also integrated ourselves tightly with the technology innovation center in the organization, which was just getting started when I last spoke to you. They’ve done a wonderful job of getting a strong team together, so now when anyone comes forward with a new idea we have this innovation center we can go to and they can very quickly put together platforms and bring in software and try stuff out, which has been a big step forward. We used to put new ideas in a queue and eventually they would become projects and it would take six to nine months to do something.

And that swings us back around to client. In the first six months I was here I did a detailed survey of all the business needs and opportunities. Client data was one of the big things, but to address that we needed a lot of these other capabilities in place, so we decided to just muddle through and address it at a later stage. Well, it is now that later stage and people are saying the biggest pain point they have is householding. How do we do householding?

If Derek is married to Denise and they have a daughter, Joni, how do I connect them? Right now all we know is Derek has some accounts and Denise has some accounts and Joni has one account. We’ve got work-arounds but they’re pretty clunky. For example, typically you would figure out that Derek and Denise are attached if they have the same address. But in high-rise apartments in Manhattan thousands of people have the same address. So we run into real constraints in some of these work-arounds and we need a lot more sophistication to be able to understand who probably constitutes a family without a lot of manual effort or going out to third-party information

providers.
What we've also recognized is that we don't just want to know that Derek and Denise have these accounts. We want to know Denise and Derek are joint owners of that account, and that Denise has power of attorney on her dad's account and has influence there. And then there are institutional advisors working with households, and we need to understand their different roles to really understand the full extent of the household.
With our work-arounds, all we know about are the primary ownership relationships. We have no easy way of viewing these joint relationships or power of attorney or any of the other relationships. So when it comes to figuring out if, say, this household fits into our private client segment, how am I going to know that if I don't have a clear understanding of the big picture?
That is what's driving the need for a client master, where we can show that Denise is a primary account holder, she's a joint account holder and she's a power of attorney, and all three of those roles reflect under her as a client. That's become the big topic of urgency in the corridors. What we're busy doing now is collecting all the user stories around this. In other words, how would we use this? So we're breaking it down into user stories which we then prioritize. This leads nicely into an agile development and implementation plan.
But to do this successfully you have to have a metadata repository, you have to have data governance, you have to have data quality tools, etc. We've put all of that in place already. If we didn't have these already we'd be looking at a very long time to value. It's an exciting time for us because we feel like we're in a perfect situation to very rapidly get value from implementing a client master.
It's most likely going to be a registry first, but over the course of a couple of years it will become the golden record that everyone will use.

And it sounds like you will realize interim value as you go.

We're looking to get real benefit from this within six-months. That's why we're using an agile approach and breaking it down into very small pieces. Each piece is going to deliver some value. Instead of releasing it into all of our client base, for example, we might just take a certain geographic region or we might classify a particular product or a particular set of securities and release that piece first and then build on and on and on.

Has upper management's expectations of what you're doing changed at all with time?

They understand it better. So many of these things take time, and some of them are more visible than others, so there was a bit of a leap of faith on their part given the extent of the complexities of what we were dealing with. I think they understand that much better now. The great thing is they're still highly supportive of the whole initiative. They see it as being critical to the company's growth.
The other great thing about the company is, a lot of my peers in the industry spend most of their time trying to prepare for the regulatory waves that are hitting the finance industry. They spend all their time just trying to get into compliance and very little time on innovation. We've been able to balance it out about 50%/50%, and I am really happy about that. When we started this journey three and a half years ago we weren't really sure how it was going to pan out. You never are. As you well know, the devil is in the details. So I think everyone is feeling good about where we are now and how we are positioned to continue to deliver real value.