DataKitchen Webinar - Actionable, Automated, Agile Data Quality Scorecards - 2025/02/19 11:46 EST - Transcript

Summary

Christopher Bergh's webinar on actionable data quality scorecards highlighted data ops principles, the challenges of data quality leadership (characterized as a "tragedy of the commons"), and the application of Dale Carnegie's influence principles to drive improvements. The webinar showcased Data Ops Test Gen, an open-source tool automating data quality testing and reporting, advocating an iterative workflow focused on specific data elements for targeted improvements. Attendees are encouraged to utilize Data Ops Test Gen and Data Ops Observability, leveraging available resources for support and implementation.

Details

- **Webinar on Actionable, Automated, and Agile Data Quality Scorecards:** Christopher Bergh presented a webinar on data quality scorecards. They discussed data ops principles, challenges of data quality leadership, and a demonstration of their company's data quality product, Data Ops Test Gen. The primary conclusion was that improving data quality requires focusing on specific, actionable items and using influence rather than authority to drive change.
- **Data Ops and Data Quality:** Bergh introduced data ops as applying agile and DevOps principles to data and analytics, emphasizing a shift from focusing on immediate problem-solving ("day one") to system optimization ("day two and three") for increased team productivity. They linked this to data quality, arguing that improving data quality and preventing production errors should be prioritized before automating deployments.
- **Challenges of Data Quality Leadership:** Bergh described data quality as a "tragedy of the commons," where shared responsibility leads to neglect. He noted the frustration of data quality leaders who have influence but limited power to

enforce change. They proposed applying Dale Carnegie's principles from "How to Win Friends and Influence People" to improve influence.

- **Applying Dale Carnegie's Principles to Data Quality:** Bergh suggested using Dale Carnegie's principles, such as avoiding criticism and focusing on others' interests, to gain influence in improving data quality. They emphasized the importance of providing actionable items, focusing on specific data elements, and iterating quickly.
- **Data Ops Test Gen Product Demonstration:** Bergh demonstrated Data Ops Test Gen, an open-source product designed to automate data quality testing and scoring. The tool profiles data, generates hygiene reviews and data quality tests, and provides data quality scores and actionable issue reports. They highlighted the ability to create multiple dashboards focused on specific data elements for different stakeholders, enabling targeted improvements.
- **Iterative Workflow and Data Quality Improvement:** Bergh advocated for an iterative workflow involving database connection, profiling, hygiene screening, test generation, score generation, issue reporting, and iterative refinement. This approach allows for quick implementation, continuous improvement, and effective communication with data owners and engineers. They emphasized the importance of focusing on specific data elements relevant to organizational goals rather than attempting to address all data quality issues at once.
- **Data Ops Observability Product:** Bergh briefly described Data Ops Observability, another open-source product that monitors both data and the tools processing it, providing a comprehensive view of the data journey and facilitating alert generation.
- **Conclusion and Call to Action:** Bergh concluded by encouraging users to try Data Ops Test Gen and Data Ops Observability, highlighting their open-source nature and the available resources such as white papers, certifications, and a Slack community. They reiterated the focus on achieving data quality through a data ops approach, emphasizing agility, iteration, and targeted improvements.

Christopher Bergh: Hello This is Chris Bergh from Data Kitchen. I'm just going to wait one more minute to start. and we will of course share slides and recordings after this call. So, we'll begin in one minute.

00:05:00

Christopher Bergh: Hello my name is Chris Bergh. I'm CEO and head chef of Data Kitchen. I'll be your host and presenter today for our webinar on actionable, automated, and agile data quality scorecards. So, we probably got about 45 minutes planned. we will, as I said, share the slides and the recording. also I will probably share a summary of it from Google in terms of text. and we'll send that out tomorrow. If you have any questions, feel free to put them in the chat window on your right. It's the little icon in Google.

Christopher Bergh: and I will periodically look at it and try to answer your questions. And of course, at the end, if there's any questions in the chat window, I'll go ahead and do it. And so you can see it in my meeting. The chat window is right here on the right hand side. And it comes from this bubble here. so if I go back to my slideshow, what are we going to do today? So we've got kind of six points to talk about. The first is the idea of data ops which is applying agile and devops and lean manufacturing principles to the production of data and analytics. we're going to talk about how to apply data ops to data quality. And then we're going to talk about the challenges of data quality and data quality leadership. And we're actually going to introduce kind of two ideas.

Christopher Bergh: I don't know if you've all read the book How to Win Friends and Influence People, but the theme of this is really presentation or influence. how do you gain influence as a data quality leader? And so we're going to talk about Dale Carnegie. We're going to talk about the tragedy of the commons an ecological term. and then we're going to actually go through how to apply some principles to gain influence and solve the tragedy of the commons problem. with data quality scorecards in an actionable automated way. And then we're going to give you a demonstration of our test gen product with a bunch of new features we've been working on that that we've released today. And then lastly, we're going to talk a little bit about how to data ops your data quality sort of how to actually go about and do it. and give you a pointer to some resources and then we'll conclude.

Christopher Bergh: So if we go back for us the purpose of our company is that we think the biggest challenge in analytics is waste and the waste comes from data and analytic projects failing poor data quality in data and analytic systems having user perceived errors people not trusting the data and just people who work with data are sort of stressed and frustrated and we think the cause of that there's probably lots of causes of it. But one way to look at it is that there's just lots of little boxes, right? you have people who data architects, they draw little boxes with lots of tools on it. Our organization structures are diverse. They're hub and smoke and mesh and centralized and decentralized. And of course, the amount of data and the things that the define tables that we create from data is just exploding.

Christopher Bergh: and so we've got lots of team boxes, and lots of data architecture boxes which create a lot of waste and problems and so yeah, I think from a frustration standpoint and as a person who've been in the data and analytics field for a long time, there's a constant frustration that the source data that you get is of poor quality and there's a frustration that we have lots of tools and that they're fragile and they break or the code that's running in those tools breaks and our customers are always asking us to do quite a lot and we're always scared of that big error that's going to show up in front of them. And so I think one of the reasons that a lot of teams suffer is they focus on day one. They focus on the immediate something is broken, I've got to fix it. There's my customers yelling at me, I got to get something done. I've got to get my task list done.

00:10:00

Christopher Bergh: they're kind of stuck in what you call Kovi quadrant one that they're always focused on the immediate tasks and I think the perspective with data ops is that you should focus on day two and day three kind of optimizing the system of people and tools and data and work and deliverables and if you do that you can actually get your team a lot more productive and it's not just us that have been singing that sort of Gartner talks about teams being 10 times more productive. We've seen it in our own use of our company and our customers and it really comes down to increasing quality and decreasing production errors and increasing development cycle time and that just means less productivity.

Christopher Bergh: And really what it comes down to is kind of before the idea of data ops, you're just spending a lot of time on wasted work, rework, doing things that don't need to be done instead of doing things for your customers. and I think that really is the focus of data ops. but how does that relate to data quality and data quality scorecard, which is the reason that we're here today? I think the first we've written a couple of books on data ops have sort of led the industry in adopting these ideas and we've quite thought quite a bit about how companies should start and I think where companies should start first is improving data quality and then stopping production errors the top of the T in this T diagram and then the lastly should be working on automating deployments

Christopher Bergh: ments and other types of automation. The sort of bottom part of the tea and luckily the way we've built our company is we've got two open-source products that really focus on data quality and these sort of monitoring production errors. One is data quality test gen and the other is data ops observability. two separate products that work together. And then we have a closed source product called data ops automation that works on that deployment and overall automation. And so what's interesting is that quality problems often start in the data itself, but they don't end up in the data. Things could break in a lot of different places.

Christopher Bergh: you could have perfect integrated data, but your customer could still see an export that's wrong or a report that's wrong or a model that's not predicting. So errors are happen everywhere along the data journey. and with teams, if we're trying to get a quality product, we have these overlapping roles like improving data quality sometimes goes to a data quality team or a data governance team. stopping production errors usually has a data engineering or an analytic engineering or an injust team. There's different terms for it. And then sort of trying to do automation is sometimes the entire data team. And so these trying to improve productivity, improve quality as overlapping roles, but for today we're really focused on the data quality team. so all our slides today really focus on someone who is focused on trying to improve the data quality itself.

Christopher Bergh: and that's what we're going to talk about today. And so what are the challenges of data quality and data quality leadership? And so I think just a definition here. Data quality is really kind of a comparison of the state of our data as it is today as what it should be, right? Is it fit for the purpose that it's going to be used to? and the challenge with data quality is the people who are in charge of storing the transactional systems or the source system, it's often good enough for them, and so data quality is a real challenge because data is a very useful product or very useful component for all other purposes in the organization. So it's good enough for them, but it may not be good enough for analytic purposes or reporting purposes.

Christopher Bergh: And so that's where this sort of mismatch comes and data quality is just a challenge for many organizations, right? There was a DBT lab survey that said data data quality is increasing. and

most people don't trust their data and there's potentially billions lost in poor data, especially with coming around the corner with Forester. And so one way to look at it is that data quality is a tragedy of the commons and a tragedy of the commons is an ecological term and you can imagine that it comes from the idea of a common shared land. So imagine that there's a shared field back in England in 1850 and one family starts to graze on it then another family starts to graze on it then a third family starts to graze on it and pretty soon the soil and the grass is completely gone. and why is that?

00:15:00

Christopher Bergh: because everyone is using this common resource, but no one is taking care of it. And so that's very similar to what happens with our source data quality. People are using it, but there's no kind of single owner. So, this shared responsibility kind of leads to neglect. And really, people are focused on short-term gains. my sheep are hungry, so I'm going to go feed them on this land that we all share. and that might hurt us long down the line, but hey, my sheep are hungry today. And as the sort of sheep sort of trudge in the soil, it creates long-term problems. And once you have this degradation, it's difficult to reverse. And so data quality is a lot of way kind of a tragedy. The commons situation and also leadership to change that to get people to improve data quality is a frustrating endeavor, right?

Christopher Bergh: we've talked to a number of data quality leaders over the years and did a survey last summer. They're sort of optimistic but fragile and they find problems in data but often are reduced to kind of data nags. you should fix this and this looks really bad. And the challenge is they need to be focused on ensuring that they're aligned with organization goals. And ultimately data quality leaders have influence but little power to cause change. And that's a really key role. And I think all of us who've been in this situation of influence but little power. and in my career as an engineer, you're trying to convince another engineer to do something. You have influence on them, but you can't tell them what to do. I think as a parent, you have teenagers, you have influence, but you can't really tell them what to do because they're just going to go do what they want. and this mismatch causes frustration.

Christopher Bergh: so do you I have learned that influence is and when a skill that I've practiced personally in my career and had to learn and so how do you get influence? there's a book that's actually written in 1936 that I think has a lot of good insight and there's this guy Dale Carboni called How to Win Friends and Influence People and it has 30 million copies sold and so it has these sort of timewarm principles on how to influence change. And so how did some of these ideas help us? Because I guess I see the data quality problem as you can only influence the tragedy of the commons.

Christopher Bergh: You can't tell people what to do. That's a rare organization that does that. But how do as a data quality leader do you get influence? let's look at a couple of principles from this book. And there's a blog I've linked here that sort of talks kind of boils it down to 30. I'm sure you can find this book in any bookstore or online. but it comes down to a couple of points that I've pulled out and then I'm going to talk about how they apply to data quality. So kind of walking through them. I think the first thing is don't criticize, condemn or complain, No one wants to hear someone who's sort of bitter and frustrated, And I think number eight is talk in the terms of another person's interest. so don't lecture someone, think of it from their perspective. And number 12, if you're wrong about something, admit it quickly and emphatically that gains people's trust. And then sort of be sympathetic with other person's ideas and desires.

Christopher Bergh: And often times in data quality, trying to improve data quality, it's not that person's full-time job. They're kind of doing it as a favor for and you've got to see it from their perspective. And

then try to appeal to nobler motives. And oftentimes what that means is focusing on specific goals or projects or programs that people have. and then use encouragement and when there's a problem, don't seek to blame, but find a way to make sure that it's easy for someone to correct. And then lastly, just make the other person happy about doing the things you suggest. and these things, I don't think of them as manipulation.

Christopher Bergh: And I really think of them as just good ways to deal with people and deal with this situation that all of us have where we're trying to influence someone when we can't actually tell them what to do. and there's a bunch more and I put the link there. but how do these apply to data quality? And we're going to go through these sort of one-on-one. So don't criticize, condemn, or complain. The first bullet. I think the challenge here is oftentimes data quality people, they're so overwhelmed with the amount of problems and the amount of work they have to do to actually identify them, document them that they're just frustrated and so we're going to talk a lot about our tool and how it uses something called generative or AI data quality to automatically identify multiple data issues.

00:20:00

Christopher Bergh: So that way it sort of takes the burden of off the data quality person instead of sort of criticizing or condemning that they've got time in the day to actually go use these influence skills instead of having in the back of their mind that they've got 10,000 lines of SQL they have to write that evening and then talk in the terms of the other person's interests. So if you're going to understand data quality, it has to be built on actionable items. It has to be built on specific limited data items. Don't boil the ocean. Don't say we've got 4,000 tables each with 200 columns. And here I can give you a list of 1.6 million data quality problems. Like that that's not actionable, right? It's not focused on what someone cares about.

Christopher Bergh: and so people care about projects they're interested in, corporate goals they have, their own specific function. And so try to focus your quality dashboards, and I use the term plural here, and we're going to talk a lot about that, of specific dashboards on specific data items, and then sort of start quickly and small and work iteratively. we really are a believer that focusing on a specific customer, building a dashboard with a customer in mind and then working quickly and learning from it is a great thing. and one of the challenges with working at iteratively is you're not spending sort of weeks or months in design. So you're going to get some things perhaps incorrect, but iterate quickly and learn from it. And then give people specific actionable items. They're busy.

Christopher Bergh: if you find problems in data, someone needs to fix it. It's usually not going to be the data quality person. Give them a package that says, " where it is. Here's exactly how you should fix it." And then, like I said before, make multiple data quality dashboards and have them align to organizational goals. And then package issues so they're easy to fix and correct. And then, of course, measure and show data and quality improvement over time. And all these things I think align to what Dale Carnegie is talking about and align to what I see the major problem in data quality teams. So I just want to check to see if there's any questions. No, good. going back to the slideshow.

Christopher Bergh: So, let's sort of walk through these and what I'm going to do is talk about our data ops test gen product and talk about some product features and then I'll go into a demo. So, this idea of we've built this open source product. we're a profitable company. we're all owned by our founders and employees. and so we've put probably five or six million dollars into building these open source products. And so why have we done that? because first of all, we think these sort of problems of observability and

data quality are just not going away. and we think that the best place to start is with individual empowerment and influence. And so, they're open- source, fullfeatured. And so, data ops, data quality test gen, it runs in your database. It doesn't require copying data. it supports about eight or nine of the top databases. It's fullfeatured. It's got an AI engine to generate. It comes with all the rules.

Christopher Bergh: and it's good for one user and one database connection. And we have an enterprise version of a very simple price model, $100 per user per month. and it's works on unlimited tables. So, two tables, 10,000 tables, based on that database connection. And it really does five tasks. it connects to your database and it profiles data and then it runs a set of what we call hygiene reviews or it generates a whole bunch of recommendations on potential problems in data and then it actually takes that profiling information and builds a bunch of what we call data quality tests.

Christopher Bergh: And how it does it looks at that profiling information, generates tests, and then every time you get a refresh of the data, it finds out if there's something changed that's significant from time one versus time two. And then we've got dozens of different automatically generated tests and about 15 custom tests that you can build based on your needs. And then what we're talking about today and what we're releasing today is this idea of data quality scoring. And so I think this first principle that Dale Carnegie talks about it's like don't condemn or criticize and I think in order to do that you have to have a tool that autogenerates both profiling information and hygiene and test characteristics. It's got to be able to do that because you just don't have time to write these tests yourselves to set up a development project to build these rules.

00:25:00

Christopher Bergh: and data is data and I think we're not going to get 100% test coverage, but it's going to give you depending upon your domain 80% coverage and the checks that you need. And we've just got different types of tests that you use. a whole bunch that come right out of the box and we learn your data, we profile your data, we automatically generate these data quality tests. So, it's really just clicking a button. And so if you've got this, I've connected to your data. I've got all this information on profiling. I've got a whole bunch of autogenerated data quality tests. that's fantastic, right? And so what we've realized is that, how do you give an individual data quality person influence? you need to be able to have them focus on specific data items. So if you look at these dots, maybe you see these dots as your data universe. And people care about very specific things.

Christopher Bergh: If there's critical data elements, maybe you're a data scientist and they have a specific model. maybe there's your top business priority, there's pick very specific data elements and score on those. Don't try to boil the whole ocean. and I think a lot of times that's where scoring fails. And maybe it's important to look for your boss to say okay we're overall data quality of our 4,000 tables your data universe but in order to make change you can't boil the ocean. You've got to have very specific scores and then work iteratively right get a process working right away.

Christopher Bergh: And so one of the things that we're going to talk about and show with our tools is ability to point it at your base, profile your data, build a dashboard, all within an hour. and that way you get a set of rules. and we believe that these getting 80% of data quality rules right, being able to start and measure and evaluate before you have standards, and try to work towards standards. So instead of doing a long analysis and thinking process kind of work quickly iterate and you end up saving a whole bunch of

time and then you can use those measurements once you've got them to kind of build your standards from data from information.

Christopher Bergh: So instead of kind of going off and working in a quiet space and you're actually having this sort of public tool that you can actually iterate and improve on these standards. And really we think that's the whole important idea of cycling and cycling quickly is and makes your customers learning so much faster. And so this principle number 19, appeal to noble motives. if you're going to have a business sponsor, right, maybe it's a VP of sales or maybe it's the CFO. they're the ones who you can appeal to saying, "The VP of sales really needs this and they need these 16 data items to be good because they've got a strategic revenue priority."

Christopher Bergh: And so when you appeal to someone who's going to fix that, maybe it's a data engineer, you can say, 'Look, I've got a quality dashboard for these specific 16 items that the VP of sales cares about for this first half of the year. Can you actually improve them? And I think that's a really important thing, Or maybe the you work in a bank and you've got 37 critical data elements and the CFO really cares about that because they're used for financial compliance. maybe it's an accounting data person or an IT person. So, you've got a specific data quality dashboard linked to that customer. Likewise, maybe you've got a machine learning engineer or she's got this model and they're really excited about it and they really want to make sure that the model runs right. So, they want to be able to look at the data quality going into that model.

Christopher Bergh: again, if the fixer is the data scientist, or the fixer could be another person in the company, have a specific dashboard. And then, not against having a sort of general purpose. Maybe the CDO wants to look at all data quality across all 3,000 tables. and that could be a quarterly review. But our point is that you need to have multiple dashboards because you have multiple customers. And to gain influence, you can't influence the entire ocean. pick specific data elements for specific customers. And then lastly, when you those dashboards have to be built on able fixable items. So our dashboards aren't some data quality score. They're actually built on a set of autogenerated or manually created data quality rules. And you can give those actionable results to people directly. And I'll show you how.

00:30:00

Christopher Bergh: and that's really key, It's not like, hey, VP of sales, they've got these 27 data elements. You should really watch them and here I got this abstract score thing. I'm looking at it and I'm saying our score is 32. this is data driven. It connects to your data and it gives the person who's going to fix that data engineer, and I'll show you they either can have access to the tool directly or we have a PDF issue report that you can mail to them that says ex you should fix exactly this and here's exactly where I found the problem. And I don't know, I've had a long time in the software industry and when you're finding bugs in software, the first thing a software person asks is, can you reproduce it? And how do you reproduce it?

Christopher Bergh: And so you have to go through this 10 minutes of work and that's a little bit of friction. So we're taking that friction away. It's a very exact way for you to deliver actionable or results. And this is what that issue report looks like. it's a PDF. s also you don't have to give it. you could send a URL. with our observability product you can also create alerts off this. But here we've just got some examples of what the issue report is. And you can see it's a very clear way even the SQL. And so any data engineer can go, I see this is how I'm going to fix it. I've got some quoted values in. Okay, I'm just going to write a line of SQL to get rid of those quoted values. And I'm going to put it as part of my injust process so we never have it

happen again. And therefore, everyone's great. you've shown them what the process is. You have the issue report. You didn't even have to type anything, right? You just had to email this thing to them.

Christopher Bergh: And then lastly, you need to show improvement over time. And so, as a data quality person, you need to show that you're awesome. And so, you need to say, we have a data quality score in January. We have a data quality score now in look at what I have done. Look at what we have helped you do and improve it. And that's why I think it's important to calculate a score and look at the longitudinal score.

Christopher Bergh: and by constantly improving your standards, you can make your score go up, changing the tests that create them, the data itself changing, you can use both those items to show how you're making an effect on the world. And so with that, we're sort of a half hour in. I'm going to stop the slides and do a demonstration. And let me just check to see if there's any questions so far. So let's look at the product.

Christopher Bergh: So this product is our data ops data quality test gen like I said this is our enterprise version open source version exactly fullfeatured you can go to our website and download it install it today it comes with some test data so you can play with it like I'm on my Mac it's running on my Mac right now it can run on your Windows machine right away it's full featured and so what does this project dashboard say first of all it's got something an idea called a table group and one of my table groups is based in a Redshift database. The other one is actually based in a Postgress database. It's called default. And so this says, okay, I've got two different databases I've connected to. And so I've done a profile of my Redshift data and I've done a profile of my data in Postgress. And on this it's got eight tables and 63 columns. And I've got something called 40 hygiene issues.

Christopher Bergh: And this is a great way to be able to quickly understand because profiling your data is just setting up a database connection. we push a bunch of SQL and we automatically create these hygiene issues. And so if I look on it and I expand this a little bit, I can see that I've also got something called a test suite. And so we autogenerate and you can manually add to this idea of a data quality test. And so a data quality test is something you can run every day, you can run every hour, you can run one once a week. And both these test suites and hygiene issues feed into this overall score of a data set. And so the second thing that we do and if we go into the data here, I want to look at our data catalog.

Christopher Bergh: And so in this catalog I've got again our default database and our Redshift database. And so I could go look at our default database. And in it it's got sort of four tables. And you can see all the tables that we have. And so if I go look at for instance the state field and the D customers, it's a Verare 50. it's probably a little bit bigger. So it's too big. we could probably get with a 40. It's a co we have what's called a semantic data type. So it's actually a code. and so we've done some distribution metrics on it here. And what's interesting is most of the people are from Texas and California and for some reason there's a lot of people from Tennessee. And so our system is automatically identified this as a critical data element just from profiling. It may be wrong, it may not.

00:35:00

Christopher Bergh: but we give you a very good first we got a little AI that identifies CDEES automatically. and so again all these things are editable by you. and we look to see if it's potential hygiene issues or test issues. So let's look at frame size in the product. And so again here is var 50. It's a code. We see a whole bunch of different frame sizes. It is a critical data element because I manually set it. But we also have a

hygiene issue. And so what that means there's non-standard blank values. So let's look at the details of this hygiene issue. So I've kind of drilled into de ebike products the frame size. And if I look at this I can kind of go in. It gives me a description of what that means. But the best way to look at it is just look at the data.

Christopher Bergh: And so here we've got a frame size, but we've got some blank values missing And this is really an annoying thing in data, should Should you not have missing data? If there is something missing, Does it say blank? Does it have a double character with a space in it? does it say error? All these different issues. And this is a way that we have what's called a hygiene issue. Maybe it's right, maybe it's wrong. And we also allow you to disposition these to say this is an issue or this is not an issue or I'm going to just deactivate in this test in the future because it's not relevant to my data set. And so all these ways are information for us in our data catalog as we go through each one.

Christopher Bergh: And so we have these 51 characteristics. we automatically develop hygiene issues. but we also develop what are called test suites. And so tests are something that are autogenerated from the data. So if I look at for instance NDI products I look at let's see I've got frame size, I've got battery life, I've got color. and so here's the case. Now I look at I've got color, red, gold. And so I profiled my data and I go look at this and I've autogenerated some tests. And if I go into this, I find out that the last time I wrote I ran the test, I've got a failed one. And I look at color has failed. And so why has it failed? And so I look at this and say, it was fine the first time it ran, but I've had some errors over time.

Christopher Bergh: Where does that anomaly come from? let's look at the source data. Maybe that can help us. And so, here we can see an issue. When I profiled the data, my baseline values were these sets of colors. And so, when I first looked at the data, now time's gone on. I've run the test. Maybe I've scheduled it. maybe as data quality person. I've scheduled it. Or maybe it was called from my data engineering team, from Airflow. Doesn't quite matter. but a new value has arrived. Orange. It's not in what we profiled. And so this is interesting, Is it not a problem? is this something you document? and in my years as a data engineer, this is always a problem because you've got groups.

Christopher Bergh: So for instance, color groups or region groups and you have a new thing show up and maybe there's another small file that aggregates or group these together and this falls through and then suddenly all the reports are off because you're missing all the orange sales and your color groups are don't include orange. And this is some kind of anomaly that we pick up and there's all sorts of different anomalies sort of row counts. and we've got 27 of these different automatically generated tests. And so why am I going into details about what this engine does? You point it at your data, profiles it automatically creates anomalies. it autogenerates data quality tests and then as the data is updates, they fire. And so all these things are very useful information for you, Because there's a whole database of really useful information that you can actually use to develop a data quality scoreboard.

Christopher Bergh: And so here I've got a couple of dashboards. And so again, what's the point? How do you get influence from a data quality dashboard? you make it very specific to one. And so here I've got my default data. I've got a CDE score. And so this is just a score just based on what are called critical data elements, the ones that we've automatically identified and the ones that are automatic. And we've got a score detail and you can actually go down and look at it by tables or by columns or by what we call semantic data types which are ways of grouping the data types together or you can actually even look at it by all our tests by the data quality dimensions. And so here if I go look at the columns I can go in and start to see the total amount is in CDE.

00:40:00

Christopher Bergh: It's got a kind of not a great score individual and it has a negative 1.22 impact on score and I can go in and actually drill in and see the specific issues that go with it. And wow, this is really important. I need to get this fixed. I can doubleclick them and then I can get an issue report. I can download that issue report. And here it comes up as a PDF. And so I can go double click on it. and I can show you I can hit double click and this should show up. And unfortunately you can't see this because it's on my desktop but the issue report shows up. And so that small issue report. so I've gone back to my quality dashboard. Let's look at another quality dashboard. So I've got one in redshift data. Now this is for a specific data science project.

Christopher Bergh: So, I've got a data scientist. They've got a model. and they only care about a few specific data elements. So, I look at this and it's got a 99.6 score. That's pretty good. But it's only interested in these three columns. and so the product photo qu qu quantities and the product ID. And there's just one issue with them. And so I can go off and look at these and say, there's a small percentage of missing values found. Maybe that's a problem.

Christopher Bergh: Maybe that's an issue for the data scientist. They assume that there's going to be data for everyone. And so again, three data elements. And so how did I get these data elements set? in my data catalog, so if I go back and look at this dashboard again, I can say that I've got my data scientific columns which are in the product name length, the product column. So if I go back to my data catalog, I look at red shift demo, I go to the products and I see the column and I look at the category name, I can see that it's been set as a critical data element or excuse me, I can say that it was set as a stakeholder group for data science. So our API this can be set manually by you.

Christopher Bergh: So I've identified this as a specific data science element and that actually drives our data quality dashboard. So this metadata that's associated with a column can actually be used to say I want my specific quality dashboard for my customer. And then likewise we can still have the general purpose dashboard. So we can go in and see one here's my Postgress data. I can go in and see it. I can look at it across all the data quality dimensions. And lastly, if I want, I can just create one on my own. So we have something called a score explorer. So I could go in and say, I'm going to go look this in my default data. I'm going to do this by CDE score. So I'm going to have my CDE score just for this. And I'm going to organize it by columns.

Christopher Bergh: So, I'm going to give it a name saying demo or webinar. And I hit add data quality dashboard. And from that time, I've got a new dashboard available for you to be able to go in and drill into and then send data items on. And kind of going back and finishing this so really our purpose here was to kind of have you to be able to have this sort of workflow in your life, to start by connecting to your database, profiling your tables, screening data for hygiene, generating data quality tests, and then when the source data is updated, the tests are executed. We generate data quality scores, and then you can review and refine them or you can share these issue reports off to data owners or data engineers.

Christopher Bergh: And so this idea of an iterative workflow, being able to connect to a generate data quality tests. This is all something that you can do very quickly inside an hour with no training. and then be able to sharing the issue details off to your team and building a dashboard. And so from a technology standpoint, if you're interested, we work and do read queries directly against these databases. it's a dockerbased application has its own database that's run in but it has its own CLI and UI that you can use

in and from a functional architecture our data ops data quality test gen kind of works against your database and it's also part of a suite of observability products.

00:45:00

Christopher Bergh: So we have our data quality test gen but we also have our data ops observability which not only monitors data but also monitors the tools acting upon data because in most cases your customers are in analytics looking at a report or an extract. They see the end result and as we stated before problems could happen in the data but problems could happen in the tools acting upon data and you need to monitor all those. And so just a quick slash to our observability product. Again this is open source as well you can work with it. And here we've got the idea of monitoring the whole data journey. That's what this picture represents. It's a data journey that is composed of observed data factory. then running some notebooks then building a PowerBI dashboard in our system collects the run status and collects all these events across all the different types of tools collecting them into one place.

Christopher Bergh: So then you can then be able to go in and build alerts and notifications and look at this kind of across all your tools and see which is running, which one had failed, which one is scheduled. and so this is a great tool to kind of monitor your entire system. And so kind of going to the last section really how to observe data quality. And I can't really emphasize enough that I look at data quality as an influence problem based on the tragedy of the commons. We've got this common data. People are overg grazing on it. It's a mess. No one wants to fix it. And it's a big field. So How do I influence change and how do I deal with all this data quality at scale? And those are really your two problems.

Christopher Bergh: And so our thinking of this is just get going quickly. focus on very specific customer needs, very specific data items and then iterate and improve and influence. And this is I think our perspective on the methodology to improve data quality. I don't think this is very different than what all the other data quality people have talked about, but this idea of agility and iteration and focus I think is a key thing. and what are the advantages of this approach? don't boil the ocean, Don't do give a data quality score of a fax number that no one uses, right? Focus on those specific data elements that are linked to someone in the organization's goals.

Christopher Bergh: VP of sales, data scientist, CFO, corporate goals, and more iteration means better standards sooner. And don't wait for data quality standards. Use a tool and start finding them and when you get something wrong, iterate them and turn them on and off and adjust them and create these quality checks automatically. don't spend months coding. and then make your trade-offs on this based on knowledge of the data, not on hope or intuition. our tool provides a great way for you to understand your data, look at it and dig into it very quickly and then tap expertise when it counts, right? So, since it has a UI, maybe you have someone subject matter experts or domain experts.

Christopher Bergh: and so it's faster and easier to implement on real data. And so being able to go in and work with real data and then really maximize your influence. that's the key. And so we've got this process that we've talked about sort of the data quality. It sort of follows our tool begins with profiling and understanding, autogenerating data quality tests, developing scores based on those tests and hygiene detectors. When you find issues, sharing them to the people who can fix it, and then use that to influence data owners and engineers to get the work done and then keep iterating and improving. Start small, work quickly, gain influence. And so I guess our goal here is really sort of data quality superpower, right?

00:50:00

Christopher Bergh: give a single person with limited time a tool to affect meaningful change. And so you can do that right now. You can download our tool as an individual, point it at data, develop a dashboard, autogenerate tasks, ship out issued reports, kind of all before snack time at lunch, and it's no cost. It runs on your laptop. and you can learn and improve over time. And it makes it easy for you to get impact and gives you these influence superpowers.

Christopher Bergh: And so if you want to learn more about these ideas and this data ops approach to data quality, we've got a white paper that you can download. and then kind of get a conclusion here. So I guess the first thing is we've got a lot of links here both to install our open source test gen and our whole observability product. We've also got a bunch of books about data quality, and we've got a seven-hour data quality and observability certification that you can take all for free. And so kind of in a summary, if you want to improve start with data quality test gen. If you want to stop production errors, add in observability and then enjoy your extra time because that's really what it's about. We're saving you time and frustration.

Christopher Bergh: And lastly, I just want to it if you start imagining could you have a tool there's a lot of talk about generative AI but really this idea could you have generative data quality that you could point at a data set it can learn your data it can screen for all these common issues and then autom automatically generate and then perform these tests and it can find the issues for you and our tool it's open source it's free and you can actually use it. So, really encourage you to and give it a try. And like I said, everything in this, presentation is going to be recorded, going to be shared, out the slides, the recording, and my voice. And so, let me see if there's any questions before we end. So, the question is, can it connect to other databases such as Oracle, DB2, or SAP?

Christopher Bergh: right now it can but it's technically possible but we don't have connections to Oracle DB2 or SAP and yeah so we encourage you to try the tool lastly just one thing if you try it out we have a health center that you can go to and look at help related for it to learn about it so we've got a full help system we also have a Slack community so you can ask questions. So, join our observability and then lastly, all the documentation that I talked about, we have all these training and certification and learning that you can get to directly from the application. so then connectors Microsoft Fabric Lakehouse. I think we support a couple varieties of SQL Server.

Christopher Bergh: So, I'm not quite sure if Fabric Lakehouse you can connect through SQL Server. I believe you can. We've had some customers talk about it, but let me check into that. And then John's got a question here. can the software compare two or more tables from different sources that are supposed to contain data looking for near duplicates that might mess up comparable totals? So yeah, our approach to that, John, is not to do sort of rowbyrow comparison of this. It's actually to be able to go in and build multiple test suites and run them against each table. And so if you've profiled a table and you have a whole suite of data quality tests and then we have a new feature in this release where you can copy those tests and point them at another table.

Christopher Bergh: Therefore, you're going to be able to instead of doing an element byelement comparison, the tests themselves are going to be able because those tests include things like row counts. They include things like all the sort of hygiene and issues that we talked about. because in my career I've never found the case where doing rowby row, column by column cases is efficient is just too inefficient. So running a whole battery of tests against two different tables we think is functionally

equivalent to doing a rowbyrow comparison of two tables. All right. And looks good. Give it a try. So follow join our Slack. We appreciate you giving it a try and thanks a lot and our team is very eager to get feedback and to have you try out our new features. looking forward to it. So thank you much and we'll be talking to you.

Meeting ended after 00:55:12 👋

This editable transcript was computer generated and might contain errors. People can also change the text after it was created.