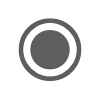
**Teams Meeting Data Deep Dive Selecting the Right Azure Data Architecture-20251028\_120720-Meeting Recording**

October 28, 2025, 10:07AM

50m 58s

 **Nicolas Blank** started transcription

 **Nicolas Blank** 0:04  
Hi everyone and welcome to Tuesday and today we are doing a lunch and learn part of the South African Azure User Group. Yesterday I quoted this interesting statistic that we were 2209 members. We have grown overnight to 2213 members I'm hoping.  
I'm hoping that one of them at least is on the call today and just for the sake of time.  
Going to talk around that today we are doing Tuesday. I should have indented that, I apologize. Today is Tuesday the 28th of October and we're doing the data deep drive, so selecting the right Azure data architecture. Ian was quite particularly with me when he says not a database architecture, it's the data architecture.

 **Ian van Niekerk** 0:50  
Thanks.

 **Nicolas Blank** 0:52  
He's going to talk about that. As of yesterday, we've had some changes, so we've created a GitHub so that all of the content is available after the fact, and you'll notice that you can stalk the speaker over LinkedIn, that you can see the folder with the recording and the transcripts.  
As well as the meetup URL in case it helps with your admin in some way. We have tried to make this as pretty as we can and we'd love any feedback. So moving on from a code of conduct point of view, this is mandatory because we are a subset of the global Azure.  
Community. And with that, in the nicest possible way, we ask you to include everyone, no matter their orientation or views. And if there's any abuse of anyone, we will just politely evict you. However, talking about people who are polite and who don't need to be evicted, I want to introduce.  
Introduce Ian Van Niekek and Ian will talk about himself. I did tell you, Ian, that I've got you because you said you had that wave, but I've got you. So Ian will introduce himself. He is one of the cleverest people that you'll meet in a hurry who knows how to talk about data.

 **Ian van Niekerk** 1:53  
Yeah.  
Um.

 **Nicolas Blank** 2:09  
And he's got a really, really interesting history that I'm not going to spoil. And with that, today Ian is going to be talking about how to select the right Azure.  
Data architecture because he does like to talk about Azure. Do you like what I did there, Ian?

 **Ian van Niekerk** 2:29  
Yeah, I see. Um, you used all of that. Uh, that clever AI stuff, yeah.

 **Nicolas Blank** 2:33  
Yes, used all of the clever AI stuff, yes, and I've reserved your likeness. So with that, I'm going to hand over to you, Ian, for you to please introduce yourself and introduce your topic.

 **Ian van Niekerk** 2:40  
Yeah.  
Yeah, good. Good afternoon everyone. Like I I briefly briefly said to you earlier, I've been in the data realm for quite a bit bit of years. I started off with some a little bit of data visualizations way back in the day.  
And then I progressively worked with pipelines, data modelling, eventually ended up as a data architect, one of the big banks and I got tired of doing it for big banks and started doing it for myself and yeah, currently.  
A consultant on creating up the data, the data capability at Cyberlogic. So today and let me quickly start sharing this side, right. So I wanted to give a little bit of context before I just go into what type.  
Of.  
What type of services that we can use for data architecture in the Azure Express, right?  
And please, this is a interactive session. Please ask questions while I'm going. It might just spoke some additional stuff, right? Yeah. So we're going to look at what is data architecture, right, and just define it briefly.  
And then why does it matter? Why do we want to look after our data architecture? Why don't we just store everything in the database and go and look for it in there? Then we're going to go into a little bit of the data blueprint or data blueprint, the typical data architecture.  
And then I'll go into a bit of a deep dive around some of the typical patterns that we find there and some of the Azure services that we can use. And then I'm going to give you a little bit of an idea where to start, what questions to ask just in general.  
um around some of the data architecture, right? Okay, cool.  
OK. So, So what, what is data architecture, right? So with the data architecture, you can think about it as a a plan or a a a blueprint of how do we consume?  
And expose data in.  
A company, right. So that's consumption of data, moving of data, storage of data, as well as the governance of data. And we then have the analytics here at the bottom, right. So you can think about the data store, the data sources.  
As the individual people, right, collecting the data across across the city, then the pipelines would be the roads and then the network moving the data from the people into the varying destinations.  
Storages is where where the people live, like the neighborhoods and like neighborhoods have got different densities and people move in and out in different places. Then we've got some governance and this this is.  
The law, right? So um.  
What do we need to keep in mind around security? How do we store in the the information? And this in a city metaphor context, that's that's the policing, right? And then lastly we've got the analytics, right? So all of these are more.  
Digital enablement, right? So technical enablement for us to get to the analytics and to get to the analytics, we need to have all of these components in there and we eventually want to answer the business question, right? So.  
It's it's putting a blueprint together to answer this type of business question. Uh, OK.  
OK, So what does it matter, right? Um.  
We don't want to start putting data on storing data and at the end of the day we realize, well, no one's using it and it's costing us a lot of lot of money and we can't use it in other facets.  
Of the business, all right. So when we start building out the data architecture, these are some of the principles that we and depending on who we chat to you, it's going to differ. But in our capability we focus on these, right? So you want the first one is for you want to.  
To build trust and reliability on the data. So it means that needs to be accurate and it needs to be where you're expecting it to be by the expected time.  
To make some business decisions from that, then also we want to make it scalable and that's where the the cloud platforms are quite good. We want to make sure that the architecture and the tech grows with the amount of data that you start consuming.  
It's not and that links into the cost effectiveness as well, right? So you will want to start with a.  
A 15 cluster server if you really just need a PC under your desk, I mean cost efficiency. So just linking into the scalability, you don't want to start big, but we also want to make sure that the storage and the processes that we use is fit.  
Purpose. If it's not, it's gonna rack up a lot of costs and and your typical like the the veracity of the the data or the how quickly the data comes in and gets stored.  
That also affects that cost cost effectiveness and that's specific. That's exactly why we want to look at the business question first, right? What business value do we wanna add? Otherwise we're gonna start ingesting all of this data.  
A real time basis and you're gonna play quite a big ball, but you're not actually answering a specific business question, right? Point #4 that we've got on here is interoperability. So I think these two go hand in hand together. We wanna make sure.  
We can start having views and use data sets across the enterprises and break down some of those silos and then collaborate effectively across teams.  
And then make sure that once we can do that, we want to use analytics and AI. In my opinion, AI is a little bit. They need to be a very good use case for AI, but in general we use analytics quite a lot in our environment and as the.  
Or it goes garbage in, garbage out, right? So we want to save it in a central repository that can get accessed by everyone, but it needs to be of good quality, right? And that's why the data architecture really matters in our case.  
OK, cool. Any questions as far, um, before we go on? Um.  
No, none. Cool. So I just, I've got to.  
Two little pictures over here, right? It all speaks to the previous ones. This is a conceptual design just to to to give over some of the the concepts and then we'll go into some of the technical.  
Design in the bottom of right. So a typical blueprint for a data architecture is this. It's gonna change a little bit depending on where you implement it and who uses it, but essentially you've got data sources.  
You want to store the data that comes in via pipelines in some shape or form. So these sources can be structured, semi structured data. This is typically Jason. Unstructured data is typically PDF documents or just free text fields that someone.  
Captured somewhere. We could also see I've got semi structured twice, so we're gonna slow it twice and then we've got some rational databases and some additional data Azure data services like.  
A I, for example, the Foundry and then Internet of Things is all of the connected stuff, right? So that's some of the typical types of data sources that we can that we can consume and then we use pipelines.  
To ingest it right into the storage layer. This storage layer is because you you've got the ingestion, but you want of events. So they they could be some of these IoT devices feeds real-time events through these pipelines.  
Don't want to store it for historical purposes or some use cases, but you also have batch processing from databases like the real time databases going into the storage layer. We I'll get you to some of the the data governance in a minute, but we.  
We would like to to apply governance as a or governance by design. So stuff like security that gets mainly applied over here. Once we we've got the data stored, we'll start processing and enriching it.  
So typical from typical example would be on a process data across data silos. The the one domains got customer transactions in the one domain.  
Mine some of the customer information and you want to typically join those together to get some real time scoring or something like that once it's been processed.  
We expose it in analytical reports. It's a little bit outdated because there's multiple ways and you'll see it here at the bottom, but the typical use case is dashboards and data visualizations.  
These days it's API endpoints that can be consumed real time. There's a multitude of words that we can get data out, but what we're really after is how do we answer a business question, right?  
So like I said in the beginning, we have to have all of this in place in this to make sure we can actually answer this, these questions in real time or not real time as effectively as possible. OK. And then there's a little block here over here.  
It is where security data management and orchestration comes in. So the security is is basically self-expandatory. So to make sure that the data that's being moved is encrypted. It's encrypted to address only people that should be able to see the data can see the data.  
Right. Not exposing sensitive data like HR data to the whole company, that type of stuff, right. Data management is just making sure that all of this information gets to it and that goes hand in hand with the orchestration it just gets.  
To where it needs to be, right? Making sure if it doesn't get to where it needs to be, it we get alerts, notifications proactively so we can start looking at that.  
Yeah, message. I'll see who ends up.

 **Matthew Levy** 15:25  
Thanks Ian. I just so maybe this is a silly question, but would you do same sort of architecture or have a blueprint for archive data or data that needs to be retained but is not part of the whole?  
Business use cases anymore. It's old data.

 **Ian van Niekerk** 15:47  
Yes.  
So, so typically how we we handle that is in the the data storage layer, right. So we'll we'll set up retention periods based on some of the use cases. So it goes through a whole life cycle where it gets used actively once it's not actively being used.  
It's quite easy with services like Azure at the moment where you set a specific time range. If it passes that time range, it actually moves it from hot data tier into a cold data tier and eventually I can't remember the tiers, but it's hot, cold and in frigid basically so.  
Depending the cost goes down, but the manager know how you retrieve it takes longer as well, right? So it typically gets the funding in this section over here.

 **Matthew Levy** 16:31  
Mm.  
Oh.

 **Ian van Niekerk** 16:43  
Oh, Mr. Blank.

 **Nicolas Blank** 16:47  
I wanted to know in this I might be running ahead, so if I am just just pull me back. What is the manifestation of this look like in terms of a blueprint? Because on the bottom I can see what it would look like in terms of a landing zone design and we can infer that landing zone design into a when there's lots of.

 **Ian van Niekerk** 16:50  
Hey.  
Yeah.

 **Nicolas Blank** 17:07  
Ways of expressing that from an infrastructure's code point of view. But if we look at the logical stuff, like if you were working and Jaco in the beginning of the call said that he's a DevOps engineer, what would you give to Jaco and what would you give to business?

 **Ian van Niekerk** 17:13  
Yes.  
Yes.  
Yeah. So what we typically give, OK, so from a typically the short answer from the business point is this layer gets exposed to to the business, right. So dashboards.

 **Nicolas Blank** 17:34  
Mhm.

 **Ian van Niekerk** 17:39  
From a DevOps perspective, this is where the data management and the orchestration comes in, right? So stuff like key vault. I don't think I stole this earlier this morning, but it's not on yet this bottom layer over here.  
Would be. This is not being depicted over here, right? But stuff in here would be keyboard, so secret management, password management, dev OPS. So are we typically?  
The DevOps engineer would sit here and he would help us deploy our pipelines right into a production environment, setting up, setting up DevOps pipelines for data pipelines, right? So pipelines for pipelines, so there's.

 **Nicolas Blank** 18:19  
OK.  
Oh, wow.  
Yeah.

 **Ian van Niekerk** 18:32  
There's a lot of DevOps in this this space, particularly setting up the environment, setting up the DevOps pipelines. How do you migrate the productionized stuff from a a dev environment into production environment, right? And this is this layer over here.

 **Nicolas Blank** 18:38  
Yeah.  
Yeah.

 **Ian van Niekerk** 18:52  
Typical stuff would be the orchestration like Airflow, DevOps, key vault management, all of that type of information around here.

 **Nicolas Blank** 19:03  
Wonderful.

 **Ian van Niekerk** 19:05  
Cool.  
Any other questions? Um.  
Oh, I'm gonna, I'm gonna just jump into an actual real like naming the the some of the some of the the stack and just bringing it back to some of the logical architecture, right? So.  
Um.  
We we used to have a, let me start at the beginning, right. So we've got in this example we've got Dynamics 365. There's a lot of non unstructured, semi structured and rational data coming on. So rational data typically is.  
Databases semi structured information like I said previously as Jason and we we we tend to get some XML files as well, right? But that then gets pushed.  
Into this this pipeline. So we connect to this this environment via pipelines. These pipelines are typically Event Hub or Data Factory or Synapse pipelines. All right, so.  
So Data Factory is more for the more batch processing type of environment where you say bring back a lot of information all at once. The Event Hub is more for streaming real time data from IoT devices, a typical.  
Um, use case that we currently have is uh.  
Vehicle assets, so of course or that's got IoT devices in there, right? So they send all of the information in here. We combine it with some of the the look up information that we get from the source systems and we store both of them in.  
Are the SQL databases or IDLS which is storage accounts right? So this is what we typically refer to landing zone or in this diagram is storage right?  
Um.  
There's a lot of theory around how you store it into the ADLS environment as well, but that's probably a whole different chat and lunch and learn on its own. It moves through various levels in here. We think to keep away from using SQL databases as a typical storage environment.  
Just because it gets too expensive too quickly. So we use the IDLS, it's relatively cheap and we can we can store a lot of information, right, which helps with the archiving as well. Matthew, it's a little bit different if you want to start archiving.  
In the the SQL database environment, um, we do have.  
SQL databases in in the storage environment as well, but we do use that as a configuration database, not as a storage container, right? OK.  
So once all of that information then lands in the the IDLS environment, you can then link it up into Fabric. So Fabric is a interesting one. So Microsoft has.  
Has consolidated and tried to make our lives a little bit easier with some.  
Fabric where it consolidates all of a bunch of a bunch of the services into a one offering. So you have one like and then on top of one like where you'll do the processing, you'll then have a SQL analytics endpoint.  
For all intents and purposes, SQL Analytics endpoint is really just a.  
SQL, you can create it as SQL, etcetera, right? So this is just a a SQL-esque environment where you can then get your information from here. So it's basically a pass through one like into here.  
Something to note about one lake is you don't only need to to link in ADLS. Relatively recent they have added shortcuts to external.  
To external vendors as well. I think you can put AWSS in here. I do not recall, I can't recall GCP also, but essentially what they're trying to do with One Lake is make it one place that you can get all of your information real time.  
Or batched or archived, right? On top of this you can then start collaborating and consuming your information. Again, this they use pipelines in here. So pipelines is basically pulling the data out of the the storage and processing environment, doing something with it.  
And then pushing it back into the one like same with the spark balls and then the analytical endpoint is where all of these like services can then start using it right. So in a real time, a bit more of a real time use case.  
You'd get a client credit score, for example, that's being fed by a bunch of telematics from a client. Yeah, it goes through the one leg and then it serves it by the API into a function so that you can get that more real time.  
It's broad use cases work similar to that. We want to have more management and reporting use cases and flow differently, right? So data factory and server in a Power BI or Power BI dashboard.  
Yahoo.

 **Jaco** 25:28  
So just out of curiosity, we what we basically do is we also use event harvest and then we use the stream analytics jobs to to get the data into the storages and then from there we use Databricks and Medallion architecture.

 **Ian van Niekerk** 25:32  
Yes.  
Yes, yeah.

 **Jaco** 25:49  
To just refine the data basically with the SQL analytics endpoint, would you suggest that would be the best way to make the basically the gold layer available to business?

 **Ian van Niekerk** 26:03  
Available. It depends on the text, the text, the tick stack, right? The later bricks is. I'm not that well versed in data bricks, but fabric is basically a.  
A a Microsoft version of the Databricks, right? So they they. So you can you can even expose it via Databricks or you can write do your computation, the box box bools in Databricks, write it back to ideal AS.

 **Jaco** 26:25  
OK.

 **Ian van Niekerk** 26:38  
And serve the data as a with a serverless endpoint as well. So there's there's a few ways you can skin a cat. There's also stuff like Synapse that you can use, right? So you can do your computations in in Databricks there.

 **Jaco** 26:55  
Yes.

 **Ian van Niekerk** 26:58  
Send it back into the IDLS world and then put a SQL Serverless on top of IDLS, which is a Synapse function, right? And that's also pretty similar to the the analytical endpoint.

 **Jaco** 27:13  
Okay, cool.

 **Ian van Niekerk** 27:14  
Yeah.

 **Jaco** 27:16  
Thanks.

 **Ian van Niekerk** 27:18  
Awesome. Any other questions?  
Happiness then. OK, so way to start, right? So there's there's been a lot of ticky ticky talk, but if the window with the rubber hits the road, there's a few questions we need to ask, right? So.  
1st is know what business problem you're trying to solve, right? It's very cool for the techies, especially to start fiddling around with tech, right? But at the end of the day, there's some value proposition that we're trying to solve for for business.  
And that's that's very important that we we first define what that business problem is right that we want to solve and it doesn't need to be you can have to take in place. I think we what I mean more about know the business problem is how do we.  
What is the shortest route to get to value first? All right, it doesn't need to be a big problem you're trying to solve, but it needs to be a business problem that we're trying to solve. Some of us are fortunate enough that there is no deck currently available, so we've green fields, but.  
More likely than not, there is a a current state in in there. So it's either on Prem that we need to replicate into the cloud or we've got some of the the the previous components in but not all. So I think it's very important that we we define our current state.  
And then make sure we we do like some steps or we define the future state, the current state and then define the intermediate steps and how we want to get there, right. But we first need to understand where we are at at the moment, right?  
And then instead of.  
This architecture right this I'm using Microsoft as a example, but this could be the architecture you can you can duplicate this architecture in AWS in GCP right? So this is instead of defining the tools that you want to use with a specific.  
Architecture set some clear principles, right? Say as a principle we don't want to consume.  
We want to do incremental loads of data, right? We don't want to consume the whole data set every day, right? That's a silly example, right? Or as a a principle, naming standards is a very basic principle that.  
Doesn't get applied constantly and becomes a big, big headache later on, right? So define the principles that you would like to implement in the data architects as part of the data architecture rather than tools.  
Tools, right? Then start small. Get a small business case, right? Don't try to solve a fraud in the whole company. Maybe try and I.  
The way clients are doing specific fraudulent transactions, right? Small data sets put the the architecture through its paces and incrementally improve, right? Which is my next point is bold iteratively, right? So.  
Keep in mind where you want to go, right? But build iteratively if if your use case is.  
Sufficient just to store data for the time being in a ADLS. That's fine, but don't know you at least need to use a ADLS account rather than starting with SQL as a base and.  
Another one of the core principles we love by in our team is design for maintenance. We can build it quite quickly with all of the AI tools and the the drag and drops, but it it's not necessarily sustainable over the long run, right? So I think one of the.  
The principles that you need to think about is how can we design A pipeline for all a architecture to be as?  
As efficient and maintenance friendly as possible over the long run. You know that's that's where I would say let's you know to start I think start by answering all of these questions that will give you good principles and guidelines on where.  
In the architecture design, you need to stop, right? Yeah. So I think we've answered a few questions throughout the whole thing. It's Kolk. What's up?

 **Schalk van Wyk** 32:23  
OK, I've got a I'm not a data expert, but got a few questions. So where would master data sort of that concept fit into the architecture?

 **Ian van Niekerk** 32:29  
Sure.  
Yeah, so you know, just go.

 **Schalk van Wyk** 32:36  
And they the cleansing. They the cleansing as well.

 **Ian van Niekerk** 32:39  
Sure. So the data cleansing is is usually part of this data, data management orchestration part, right. So that happens throughout the process and the medallion architecture that Jaco referred to earlier.  
That's also part of it, but the short answer is it happens in multiple stages throughout. So cleaning your data, you'll do the technique what we call technical data cleansing before we store it, right? So removing some of the null the the empty values.  
Making sure it's costed as the correct data types. We'll do this in this step, right? Business rules and business cleaning will happen in the the processing and enrichment phase, right? Master data management depends on who you talk to, but.

 **Schalk van Wyk** 33:20  
OK.  
Exactly.

 **Ian van Niekerk** 33:33  
My, my.  
My opinion is it needs to be part of the source systems, right? This the the master data management basically sits over this and and this is one of the feedbacks into the master data management. Master data management is then fixed at source and.  
Then it comes in here again, right. So it's a little bit of a a feedback loop in here. I'm just making some notes. I I like the questions. I'm gonna, you know, answer my pranky with it, you know?

 **Schalk van Wyk** 34:02  
Yep.  
So and then obviously that fits in with your data from operational systems coming to your master data and managing it through that. Yeah, OK. And then I mean.

 **Ian van Niekerk** 34:19  
Yes.

 **Schalk van Wyk** 34:23  
It's quite difficult to have that most data management is a whole story by itself, I would assume.

 **Ian van Niekerk** 34:31  
Yes, so, so, so that's why I say it depends. That's an ideal world, right? We it's a separate system. Practically what happens in the field is they've got some type of master data management running inside of the the data storage, right?

 **Schalk van Wyk** 34:49  
Mm.  
OK.

 **Ian van Niekerk** 34:52  
So it comes in here, it looks at this. I'm gonna say look up for lack of a better word and then cleans it as it goes into the storage, right? But that still means the data in the the source environment is still dirty, right?

 **Schalk van Wyk** 35:00  
Hmm.  
Yes, exactly now.

 **Ian van Niekerk** 35:06  
So that's why I say it's a little bit more practical to take it through here and then store it as a clean version. But to solve the problem holistically, you need to fix it at source, and that's why it says.

 **Schalk van Wyk** 35:14  
Yeah.  
Because one system might affect another system, so you have data flowing so it becomes a bit of a chicken egg situation. And then transactional data. I mean if if you ingest transactional data, would it be?

 **Ian van Niekerk** 35:21  
Correct.  
Yeah, exactly. Exactly.  
Um.

 **Schalk van Wyk** 35:35  
Be best to store it sort of raw and then only process parts out of that, yeah.

 **Ian van Niekerk** 35:40  
Yes. So that, yeah. So that's exactly where the the the medallion architecture comes in, right. So it might be worth while touching on it, just maybe briefly. So we've got 33 levels right of medallion architecture. It's the bronze, silver and the gold.

 **Schalk van Wyk** 35:50  
OK.

 **Ian van Niekerk** 36:00  
That all sits in this data storage layer, right? So bronze is a one to one replication of our source system, right? Going from bronze to our silver environment, we start doing some technical cleanups.  
And then start aligning it a little bit more to the what the business, the business use case is at the end of the day, right. And then when we move it into gold, we do a lot of the transformation and enrichments and that then looks almost exactly like to answer. So we move it from an OLTP system.  
In the OLAP system from bronze to to gold, right and that then aligns with the. So on the bronze layer it's a lot more system or source aligned where on the gold side it's a lot more business, business use case aligned.

 **Schalk van Wyk** 36:42  
Thank you.  
The last question, so different you you can't replicate this over environment obviously. So this will be your production environment unless you have some big, some big company like the old mutuals or something probably.

 **Ian van Niekerk** 37:07  
So.  
Yeah, yeah, yeah. So, but so, so it depends, right. So we've got small companies that run the same thing, but multiples of things, right. We you can have multiples of things. What we found though is.

 **Schalk van Wyk** 37:21  
Mhm.

 **Ian van Niekerk** 37:29  
Typically the users want to test on production environment, right? So what typically happens is the storage and the the pipeline environments. We have a dev and a production environment, so pipelines and data data sets.

 **Schalk van Wyk** 37:43  
Mhm.

 **Ian van Niekerk** 37:45  
Where we actually process the data, you have one big pool, but it writes back either to the product production or dev environment and sometimes the the company split up the analytical and reporting sites, right? But because it's based on production in source.

 **Schalk van Wyk** 38:00  
Mm.

 **Ian van Niekerk** 38:03  
Uh, data. They they tend to use the environment, right? Um, you know?

 **Schalk van Wyk** 38:10  
Cool. Thanks.

 **Ian van Niekerk** 38:12  
Oh, Jaco, I saw your hand was up briefly as well. Do you still have a question or?

 **Jaco** 38:18  
No, my comment was just towards the validation of the data. That is what we use our stream analytics job for. So as we pick up the events from Event Hub, we first validate it and then it goes to storage.

 **Ian van Niekerk** 38:28  
OK.  
OK, what what type of validations do you apply on on that? Is it like technical validations to make sure it's in or would like?

 **Jaco** 38:41  
Yeah, it's it's basic, sorry, yeah, it's it's basically like contracts just to ensure that everything that is needed is is provided.

 **Ian van Niekerk** 38:47  
Mhm.  
OK, that's pretty cool. I think that's a good a good use case for for that. And do you guys use that data that you pump through the Event Hub in real time as well or is the the Event Hub just used for?  
Uh, the ingestion port.

 **Jaco** 39:11  
No, it's basically just for for for analytics, so it's after the fact and not real time.

 **Ian van Niekerk** 39:17  
OK, cool.  
Any other questions? Oh, sorry, last call.

 **Schalk van Wyk** 39:21  
So so that's also that's also we so getting data from external systems that data factory doesn't fit because it's too expensive and it's for running continuous like if you have instead of and also event hubs won't necessarily work for.

 **Ian van Niekerk** 39:34  
Yeah.

 **Schalk van Wyk** 39:40  
We have to feed data from external systems or ingest it to actually feed it in. So we're using Azure Functions and whatever to try and pull that data in.

 **Ian van Niekerk** 39:44  
Yeah. So do you? Yeah, yes. So, so, so they this, there's a few stuff here, right. So we initially started with data factories running the pipelines and to your point, it was too expensive, right.  
Then we said OK, well we need to use something different. So we actually was in the the ingestion phase. We're using sparkles to pull in data. So we we technically using.  
We're still using Data Factory, but it's more for orchestration of the spark pools, right? Then actually using the pipelines and the data flows for the computation, right? That that brought down the cost a lot.  
Then we said, well, we could possibly take down the cost even further and that's where we started using logic apps. We still use pipelines as an orchestration method and we've got a database that that sits over here which is configuration based, right. So you say I want to bring in AB and C and that.  
Then goes the Data Factory loads the the config information from here and it executes the other logic apps, spark pools or pipelines depending on where we're getting the information from and that that we've seen like.  
That that's the cheapest that we can get away with a a serverless environment.

 **Schalk van Wyk** 41:16  
OK, so it's a viable. If someone says why are you doing it like this? It is actually a viable option because I've done logic apps, Azure functions and then Azure durable functions because that does orchestration instead of having data factory.

 **Ian van Niekerk** 41:22  
Yeah.  
Yeah.  
Yes. So and and the other thing is the other reason why we're using functions and Spark pools is or Pi Spark is it's a lot more flexible than the the pipelines that they that that's in there, right. So especially when it comes to pulling data from.

 **Schalk van Wyk** 41:48  
Um.

 **Ian van Niekerk** 41:51  
We found that that's very difficult to. It's possible, but because API's authentications is so varied, we use a function app. It just makes it a lot easier to bring in the data with that, but we still use the Data Factory as an orchestration layer.

 **Schalk van Wyk** 41:58  
Pentication. Mm-hmm.  
Yep.  
Yeah.  
Yeah, you have one generic process that goes and says this is what I extract, transform, load and then it'll call another function to go do authentication or whatever it might be. OK, OK.

 **Ian van Niekerk** 42:13  
Yep.  
Exactly. Um, you know.

 **Schalk van Wyk** 42:25  
Um.  
And what about using Python? Sort of. Where would that if you had a Python, your own custom Python?

 **Ian van Niekerk** 42:34  
Yeah, so, so, so Python was is in the spark pools, so you so the the the spark pools can sit in architecturally. It's not really a part of.

 **Schalk van Wyk** 42:38  
Smock pools, OK.

 **Ian van Niekerk** 42:48  
Either fabric or data factory, right? It technically sits over here. So you can refer to the spark pools. You spin up a cluster that sits over here. I think the minimum is 3 nodes in by about $2.00 an hour of processing, right? But.  
You then in the Spark pools you use Spark. You refer to the data sets, you bring the data sets in from ADLS. You do your computation, you read it either you write it back in here or you store it in the one lake and we still use the data fabric or synapse pipeline.  
To do the orchestration, decide which which Python notebooks we need to execute and when it needs to execute.

 **Schalk van Wyk** 43:36  
OK, I see there's a new data wrangler for Visual Studio code.

 **Ian van Niekerk** 43:48  
Oh, I haven't seen that yet. Yeah, no.

 **Schalk van Wyk** 43:48  
Yeah. So I just that's that's a question sort of that. That's why I was wondering. OK, don't worry.

 **Ian van Niekerk** 43:53  
Yeah, you can ask a question, but I I don't promise I'm going to have an answer. Yeah, yeah, I haven't seen it yet though.

 **Schalk van Wyk** 43:58  
I just saw it recently, so that's why I'm asking.

 **Ian van Niekerk** 44:05  
Oh, any other questions?  
Oh, if you want to, guys want to. I'm just going to use my little mouse and Mr. Blank, do you have some some closing comment? Yeah.

 **Nicolas Blank** 44:19  
I just want to let you know that your 10 minute warning is here.

 **Ian van Niekerk** 44:23  
OK, I'm finished. So if there's no other questions, there's a QR code, yeah.

 **Schalk van Wyk** 44:28  
But there is the data, the data, what's it now? The fabric. Yeah, fabric. So I mean, I went to the presentations the other day about it, but there's very vague really where that's going.

 **Ian van Niekerk** 44:38  
Yes, this.  
Yeah.

 **Schalk van Wyk** 44:48  
I'm not sure. Does one sell it to your business? This is like down a line stores. It seems to me it's like version one almost.

 **Ian van Niekerk** 44:54  
So.  
Yes. So what Microsoft did with Fabric is they did a very, I don't want to say very quick, but they did a MVP, right. So they they have a minimal viable product and they released it. They're working with the community basically to improve it as they go along.

 **Schalk van Wyk** 45:06  
Mm.

 **Ian van Niekerk** 45:16  
Right. So the the downside with that is you don't have all of the functions that you would have had with Data Factory or Synapse from day one, right? So but the the other side is you can start using fabric a lot quicker.  
And have input in in getting a day, right? Um.  
That being said, is public is definitely the way Microsoft will be going. The the advice that we give our clients at the moment and it's changing a little bit is stick to what you stick to Synapse because we usually go with Synapse, but stick to Synapse and literally reevaluate it in three months, right?  
Some of the clients is in some like if you asked me this question six months back, I would have said listen, let's stick with Synapse for now. If you ask it to me, I'm going to say it depends if you use some of the the more, the more technical or the more advanced functionalities.  
Fabric is a little bit lacking on some of the stuff, but it's getting there. It's a lot better than six months ago, right? So we are moving some of our clients over from Synapse into the data, not the data factory, the the fabric capacities.

 **Schalk van Wyk** 46:34  
So put it on the technology radar.

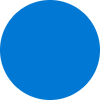
 **Ian van Niekerk** 46:36  
Yes, exactly.

 **Schalk van Wyk** 46:37  
Oh, thank you.

 **Ian van Niekerk** 46:41  
Oh, Bing.

 **Dean Sardinus** 46:43  
How's it guys? Sorry, can you hear me? I just want to check.

 **Ian van Niekerk** 46:44  
Hey, yeah, yeah, can we hear you?

46:45  
Sure.

 **Dean Sardinus** 46:47  
Um, so just on the on the fabric topic, um.  
What what are the your you know ingestion options that you have with in fabric you know so is it data gateway?  
That is the main tool that you use and and what about if you know you have customers where they don't want stuff to flow over the Internet, maybe they've got express route or large, large data. Now how do you how do you get the traffic or the data?

 **Schalk van Wyk** 47:08  
Yeah.

 **Ian van Niekerk** 47:19  
Yes.

 **Dean Sardinus** 47:22  
Cross.

 **Ian van Niekerk** 47:22  
So so so fabric is you can it's basically they they took data factory and migrated into data fabric. So the way you get data into fabric is basically the same ways that you you get the data into factory IDF.  
Right. But that being said, is it depends on the way.  
That the client prefers, right? You can have a very secure VPN gated gateway or tunnel and still connected to fabric. It's a lot of effort and it's very technical, right? So it's a lot of effort to get that up.

 **Dean Sardinus** 48:05  
Yeah.

 **Ian van Niekerk** 48:06  
It's VPN gateways, but essentially fabric can just see that one database with a tunnel via the VPN, right? Alternatively, what you can have is a gateway server.

 **Dean Sardinus** 48:16  
OK.

 **Ian van Niekerk** 48:23  
In the client environment, they call it a SHIR runtime integration runtime. Can't remember what there is an SHIR host integration server hosted, self hosted, self hosted integration run so that you use the.

 **Dean Sardinus** 48:32  
Self self hosted drunken.  
Yeah.

 **Ian van Niekerk** 48:41  
SHIR on a server. That server is essentially the gateway into the whole network, right? The other thing is what we usually do is we use SHIR because all of the traffic between the factory and the SHIR server is encrypted, right?

 **Dean Sardinus** 48:48  
OK.

 **Ian van Niekerk** 49:00  
And then we just give the SHIR server visibility to the data that we want to see, right? So we don't expose it to the whole network, we just expose it to parts of the network that we want to see. And we haven't had any problems thus far with it, but it also depends on how security you really want to go.

 **Dean Sardinus** 49:08  
Yes.  
OK.

 **Ian van Niekerk** 49:20  
So you can make it very, very complex or you can keep it simple.

 **Dean Sardinus** 49:24  
All right, so the integration runtime, obviously it's sort of a VM sitting privately. So as long as favorite solution can see the public IP address, then it should be able to connect OK.

 **Ian van Niekerk** 49:35  
Yes.  
Correct. Yeah.

 **Dean Sardinus** 49:40  
Thanks.

 **Ian van Niekerk** 49:44  
Oh.  
Any other questions?  
So there's a QR code for you guys if you want to connect on LinkedIn or just reach out. But I do think Nick also put my my e-mail address on on his side as well. So if you guys, if you have any other questions, please feel free to give me a shout and we can correspond accordingly, OK?

 **Nicolas Blank** 50:05  
I did.  
That's amazing, Ian. Thank you so much and thank you so much for for being willing to share and to to give your knowledge away. So thank you very much.

 **Ian van Niekerk** 50:19  
Sure.  
Yeah, big, big pleasure.

 **Nicolas Blank** 50:26  
All right. So with that, we're going to give everybody back 3 minutes of the day and we thank you for joining us on today's lunch and learn. And as I said earlier, the session is being recorded and Ian, I expect there to be a long tale of.

 **Ian van Niekerk** 50:27  
I see.

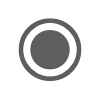
 **Nicolas Blank** 50:43  
Folks watching the recording, so watch out for that and you may even get questions over the next year.

 **Ian van Niekerk** 50:49  
Oh, happiness. Happy to help. Thanks guys. Have a good one.

 **Nicolas Blank** 50:53  
Thank you everyone and bye bye. Cheers, bye.

 **Dean Sardinus** 50:53  
Thanks.

 **Jaco** 50:53  
Thank you. Great. Cheers.  
Thanks guys. Bye bye.

 **Nicolas Blank** stopped transcription