The Measure Pod: #43 What is data retention in GA?

[00:00:00] **Dara:** Hello, and welcome to The Measure Pod episode number 43. This is a place for people in the analytics world to talk about all things analytics and data. I'm Dara, MD at Measurelab.

[00:00:26] **Daniel:** And I'm Dan, I'm an analytics consultant and trainer also at Measurelab. Just a couple of updates in terms of news items that have happened in the last week or so. Not super exciting, but it's worth mentioning a couple of new releases in Data Studio this week. They have had a bit of a rearranging of their like admin editing panel on the right hand side, when you're editing a dashboard and they've now got something new or they've at least renamed one of the panels called the data panel. A little quality of life improvement, but it means that you've got access to all of the data sources and all of the dimensions and metrics all within one interface. You don't have to switch out data sources to then access the new dimensions and metrics. So it's a small thing that's made building dashboards that little bit easier.

[00:01:04] **Daniel:** Data studio have also released something called a Linking API, or at least they've released it into general availability. And what this API is, is a way of creating dashboards on the fly really easily, rather than having a single dashboard that people then have to copy paste and manually replace all the data sources, which could be a bit tedious or time consuming. There's an API to kind of generate dashboards on the fly, which is pretty interesting. I've not had a play with it myself yet, but I'm looking forward to having a little mess around at some point in the future.

[00:01:30] **Dara:** I was gearing up to pretend to be excited about those updates, but that GDS Linking API, that actually is a potentially interesting feature.

[00:01:38] **Daniel:** Yeah, it can be, this is the interesting thing about this update is that it's like a new thing. It's like a new way of creating dashboards via, rather than creating one and sharing some public URL and swapping everything out. This can be used internally, but as well as externally, people can share templates via this API URL, someone clicks a link and then you've instantly got it accessed to your data and you can just save a version if you like. I think just on the side, it might actually mean that a lot of people won't actually have to save dashboards anymore. So if you can just click this link and every single time view the same dashboard with your data, why would you ever have to then click edit and save? You never have to save a version of something, you can just dynamically create a dashboard on the fly for you whenever you need it. So it's a really interesting concept rather than creating, saving, and editing dashboards, you can just in a way view a temporary dashboard whenever you need. I'm looking forward to the applications people come up with.

[00:02:28] **Dara:** Yeah, no, likewise. Okay, but moving on from the news we've got a good topic this week. We've got one that I think is around a feature that maybe sometimes, or a setting that slips off the radar if that's a phrase? Skips the radar, slips through the net, which is data retention. And this actually came from a conversation we had internally at Measurelab where the data retention settings were actually affecting some reporting that we were trying to do. So our topic is what is data retention in GA (Google Analytics)?

[00:02:55] **Daniel:** Yeah and as you said, it came up recently, internally when we were talking about this, because it affects not just GA4, but Universal Analytics as well, and quite often, and I've spent a lot of my last couple of months training Google Analytics 4 to lots of people. And whenever we get onto the concept of data retention, whenever we talk about data retention and as we go through the admin settings, it's always a bit of a surprise and a shock of like, oh, I can't believe Google are doing this and then I'm like, they do this in Universal Analytics by the way. And they're like, oh, I didn't realize they weren't keeping all of my data forever. It's a couple of things, it can catch us out if we're pulling bigger reports or longer reports in Universal Analytics or GA4, but there are some differences.

[00:03:31] **Daniel:** So I think with the focus on GA4 over the last couple of months, and you know, only ever increasing in the next couple more, let's talk about data retention but it's not new, this isn't anything different to what we've had before, but there are some small, subtle differences between the two that I think we can explore.

[00:03:46] **Dara:** Okay well, let's just maybe set the scene a little bit. I'll start right at the beginning, data retention is a setting as you rightly said, it's in GA4 and it's in Universal Analytics and it's to control how long user and event level data is stored in the Google Analytics servers.

[00:04:03] **Daniel:** That is the nuts and bolts, it is a way of Google not keeping data forever because data storage and querying and processing, if anyone's used anything like BigQuery, because they do use BigQuery behind the scene. It's expensive, or it can be expensive, especially for the bigger sites and apps. So Google, they're not doing this at the kindness of their heart, they've got an angle, like we probably mentioned a hundred times on this show, Dara. The angle is about advertising and monetising you in different ways, even if it's not via GA (Google Analytics) directly, but because storage of very granular level data can be expensive, they don't do it. I think that's the long and short of it.

[00:04:34] **Daniel:** So what they've got is this feature called data retention and the way you can access that in GA4 is under the property settings. There's a section called data settings and then data retention, and then in Universal Analytics it's very similar, it's in the property settings again, under tracking info and again, data retention. As you said, I'm just going to say the same thing so slightly longer but it's just a way for data to be kept for a certain period of time before it is removed from that raw data tables in the back end of Google Analytics, which again is a BigQuery database.

[00:05:02] **Dara:** And the default for GA4 is two months and I'm a bit sketchy on what the default was for Universal Analytics, which I should know because Universal Analytics was my specialist subject, but past tense I've forgotten. Do you remember, was it 14 months?

[00:05:17] **Daniel:** I think it was 26, but again, don't, don't hold me to that. And yeah, don't write in if I've got that wrong.

[00:05:23] **Dara:** It's all about GA4 now anyway.

[00:05:24] **Daniel:** It's all about GA4 exactly. So this is actually why it's become part of the conversation more recently with GA4 and actually the feature itself is not new. What is new for GA4 is that window, the default window, should we say is actually a lot shorter. So in Universal Analytics, even on the free accounts, you have various options. We'll go through and explain exactly what this feature does because it's not as terrifying or as daunting as it might first seem but in Universal Analytics, the options you have available to you are 14 months, 26 months, 38 months, 50 months, and there's even a setting called do not automatically expire. So in a sense, there is an option for Google Analytics to never delete your raw data, but that's not the default. The default we think is 26 months. However, in GA4, there is only two options in the free version of Google Analytics 4 and that is 2 months and 14 months, and the default now is two months.

[00:06:15] **Daniel:** So there's a big difference, there's a big difference from even having an option to say, do not expire data. So do not delete data to, we are going to keep it for two months. So it's a big old step change. So in a sense, what it might do is a lot of people might not even be aware that they have a data retention setting and they might even be set to the 26 months. But I feel like being set to two months, you are more likely than not going to find that out sooner in GA4 than you ever would in Universal Analytics. So I think it's probably about time to explain what this actually does and how it works and how you might not have ever realised that you've got this and still using your historical data absolutely fine in Google Analytics.

[00:06:51] **Dara:** It's entirely possible you won't have even noticed the effects of this setting in Universal Analytics in the past. It doesn't apply to all the data, so all of your kind of pre-aggregated data that's stored every day in GA (Google Analytics) that's unaffected. So if you look at all the standard reports within GA (Google Analytics), just like you can view all of that data without the risk of sampling, same applies with the data retention settings. In Universal Analytics at least it's really if you are applying a custom segment, or if you've got a particularly advanced custom report that then this might show up some issues in the data. And it's basically where the user or event level data associated with an identifier. So it could be a cookie or it could be a user ID, or if it's an app, it might be a, an advertising ID and that data wouldn't be kept beyond the retention period.

[00:07:39] **Dara:** So from the previous interaction of that user, if you've got your retention period set to 14 months, then after 14 months of inactivity that user and event data would no longer be stored on the Google Analytics servers.

[00:07:53] **Daniel:** I think that's a really interesting and really big distinction or at least when we speak to people is because people often don't venture outside of those standard reports. And when we talk about standard reports in Google Analytics it's just anything on the left-hand navigation, if you click into the source medium report, the pages report, the channel report, these are all standard reports that for one, as you said, won't be affected by sampling, which is great. But secondly, won't be affected by this data retention period, because they're in a sense they're pre aggregated. They're daily totals, right? There's nothing raw data-y about them at all. They're just pre-calculated tables that are a lot easier and cheaper, more importantly to store and to process. So, as you said, when you're using things like segments or secondary dimensions, or even creating custom reports, that's when this thing kicks in, when sampling kicks in and data retention occurs and exactly the same in GA4 as well.

[00:08:38] **Daniel:** So in GA4, same approach, different UI, I suppose the reports workspace that won't be sampled, but there is no data retention affecting of that data. So if you log into GA4, where you land in the reports workspace, that data will continue to be there. Daily users and sessions and events per day that's not going anywhere. Where it does affect the GA4 data is in the things like the explore workspace when you're creating segments or you're doing custom reports or tables or visualisations over historical data. So this is really where we are going to be limited because we have to go back to the raw tables to process the vis or the table that we need and that's where the data just doesn't exist.

[00:09:16] **Daniel:** So the data retention period, you might not have even been aware. A lot of people I speak to aren't aware that it even exists because you might have stuck quite nicely to those standard reports and actually you haven't ventured out of that ever or not yet and in which case you are probably not putting up against this. But in GA4 I think more people will because when we come to look at the reports in GA4, quite a lot of the interface is quite minimal, right? And I know you can create your own reports in the interface that you'll use in the library, but actually quite a lot of the reporting, the analysis and the explorations we'll be doing will be done in the exploration workspace and that's really where we're going to find out if we've got only two months of data or not.

[00:09:50] **Dara:** Might be a good point to mention as well that there's other, for example some of the demographics data will only be stored for 6 months in Universal Analytics, and I think it's 2 months in GA4 and that's regardless of what data retention settings you have, although if you set your data retention settings to lower than that amount then it will respect the data retention settings, but it won't store it for more than that amount no matter what you've set your data retention to. And I think Google Signals as well is 26 months and I might be wrong, but I think that's for Universal Analytics or GA4, it will only keep it for a maximum of 24 months.

[00:10:26] **Daniel:** All these things are quite nuanced, and I think this is the interesting part is that Google Analytics got the perception of it can be that this is a free tool that will store all data for me forever. I suppose the kind of big shake up we've come across in the last couple of months is the fact that they're going to be turning off Universal Analytics as of the first, next year. And then they said, then you'll have access to the data for six months and then after six months, there's no guarantee that you even have access to the data. So in a sense, what Google have said is that potentially, and there's no guarantee of this right now because I suppose it's so far out, they haven't confirmed. Potentially as of the 1st of January 2024, all of that historical data, all of that previous data you had in Universal Analytics, all those tables, aggregated and non-aggregated are probably going to be deleted. That's basically what they're saying is at some point in the future, they're going to be deleting them.

[00:11:08] **Daniel:** So I think before we've just kind of thought this is a safe space to keep all this data, haven't really thought about it too much, but actually I think the interesting part is the kind of Signals data and the other bits of data as well. I mean, even if you look at things like the Search Console integration, some of the reporting you get in there, there's always a bit of a lag and it only keeps the last 90 days I think. There's all of these different data sets in Google Analytics, and I think this is really proving the point or justifying the idea of like owning it yourself, and that's basically what we're describing as a data warehouse. Pull the data out and keep it yourself, keep a record of it for yourself and then you get to control the retention, right? You get to control what you keep and where, and for how long.

[00:11:41] **Daniel:** And actually this is possibly one of the reasons why they release the BigQuery connector for free in the GA4 properties, because not only could they say, if you want to keep the data, you can forever that's up to you, but you pay for it. In a sense, they've just flipped that on its head and they said, okay, we're not going to pay for this anymore because it can be quite expensive. So you can keep it, you just have to start funding it. And I think this is the interesting approach, and I think this is why they've gone down the two-month data retention period because they've given the BigQuery connector out. And this is also probably one of the many times we've said this as well, Dara, but like always get the BigQuery connector set up as soon as you can, even if you have no plans on using it right now, because your future self might thank you for doing it when you did because in let's say a year or 18 months down the line you decide to use it, you'll wish you had the historical data there. It's all about keeping historical data to do what you want with it down the line. There's only one way of doing that now and that's via the BigQuery export.

[00:12:31] **Dara:** You just made me think of something when you're talking about historical data as well as another difference between Universal Analytics and GA4 in terms of the data retention is, if you change the setting with Universal Analytics, it didn't apply that change to data that was already collected. Whereas with GA4, it does. Still that won't make a great deal of difference to you, but at least it is a benefit, it means if you increase your data retention, say it was set to two months, which is the default and you suddenly realise, and you want to increase it to 14 months, that will apply to any data that you have already collected, which is a good thing.

[00:13:01] **Daniel:** Yeah, for sure. It's also not something you probably wouldn't be changing that often, to see the effects of. It's also worth noting as we're talking about the setting itself, there is a, a slider underneath, there's a secondary option with this, not just the window itself, but there is another option within the window to reset user data based on new activity. So this is actually on, by default for GA4 properties. Again, bit more fuzzy on the Universal Analytics side, but the way to think about your data retention is with this enabled is not like a tidal wave kind of deleting data with a two month lag, right? So it's not like the rolling two month period is only ever kept. It's keeping data, all data for all users forever until that user has been inactive for more than that period of time. So in a sense, what this means having the two month data retention, which is the default and having this reset user data on new activity enabled, which again is the default. What that means is that all user data will be kept forever assuming they visit your website or app every month, assuming there's some kind of activity for them within two months.

[00:14:02] **Daniel:** As soon as they've stopped visiting the website or the app for 2 months, then their historical data is deleted. So it is not a perfect, it's not easy to say exactly when you have perfect data from or full data from. So I always treat it as 2 months anyway. Because before two months, you've only got users that have been on the website within the last 2 months. It's a very odd way of thinking about things. So in a sense I do talk and explain when we are talking to our clients and other people about this setting, I do explain that this is the case though, this reset user data on new activity is enabled. And it means that, you know, you keep all your data until they don't visit for 2 months. But because it's so ambiguous beyond two months or beyond that window, I treat it the same regardless. So I still think of it as a 2 month window or a 14 month window or a 50 month window, because pre then I have no idea how much of the data actually exists anymore. So yeah, I'm not sure the actual value of that setting because it doesn't change the way I use the data, but again, that's just me.

[00:14:56] **Dara:** Yeah I noticed something weird when I was reminding myself of some of these limitations and some of the details around this setting. There's something in the documentation around that toggle that allows you to reset on new activity. And it says that it only applies to user level data, which I didn't fully understand. If it's saying it doesn't apply to event level data, it seems strange to me that it would apply to user level, but not event level data. You familiar with that? Do you know why that is?

[00:15:23] **Daniel:** No, is that GA4 and Universal Analytics?

[00:15:25] **Dara:** It doesn't well, you know, classic documentation. Because for some of this support article, it does specify whether it's Universal Analytics or GA4, this particular little single sentence of doom doesn't say whether it's one or the other or both. So I don't know, yeah. I was just a bit thrown to see that. I don't know why it would apply to user level, but not event level.

[00:15:45] **Daniel:** I'm wondering if it is just another standard, Google documentation which is that it's just poorly worded and that it's actually all data, but they're just saying user data, because you know, everything happens from a user, right.

[00:15:54] **Dara:** Potentially, yeah you might be right. It just might be the wording. So on a kind of similar thread, you mentioned earlier about the do not auto expire option. We just talked about the reset on new activity, which is a way in theory at least to never delete that data. Although, as you said, we're not entirely certain about some of the issues around that, but this do not auto expire option is actually on something that was only available with Universal Analytics it's not available for GA4, including the 360 version, but that was an option available to you to, I say, was it still is if you're still using Universal Analytics, So as long as that's still alive, you can use the setting to not auto expire the user and event data, but that option's obviously going to go away and maybe we'll be wrong, but I don't think that's going to be introduced to GA4, that would be my guess.

[00:16:45] **Daniel:** One would imagine that it won't be, especially as they've now provided the BigQuery export, but also if they would've done, they would've done it by now I think. Interestingly, differently to Universal Analytics, the GA4 options are different if you're on the 360 version and the free version. So on the free version you've got 2 and 14 months is your two windows. But on the 360 version, you've got 2, 14, 26, 38 and 50. So you can extend that window to do, you know, the kind of custom reporting if you're paying for it. So again, I think that's another aspect of this, but again, they've not got the do not automatically expire even on the 360 version of the product. You would've thought they would've put it in, if they were going to now, I think, I feel like this is more of an active decision to take it out rather than them not thinking about it, you know?

[00:17:26] **Dara:** Yeah, exactly. I would imagine it was in there originally when this was introduced to Universal Analytics because of the fact that it was, you know, it was a change to introduce retention settings. So they probably originally left the option to override it and then with GA4 different, different ball game, and they can just start without it.

[00:17:45] **Daniel:** We've mentioned it before the podcast as well, Dara, the reason why it's become such a hot topic or a subject of kind of focus of conversation recently is because of the idea of exporting your data from Universal Analytics as of next year. So as of July 1st next year, it stops processing new data, you get three month grace period, if you're on 360, but it still stops at some point. And the idea then would be to export your data from Universal Analytics, all your historical data that you want to keep, pull that into a wherever, you know, a notepad file or even a warehouse somewhere, or a BigQuery database, you know, whatever you want to pull into so that you have access to that yourself. You are paying for it, your storing it, you are keeping it, but you have access to it.

[00:18:21] **Daniel:** The thing that a lot of people don't realise is that you won't be able to pull out the raw data if you don't have the data there to begin with. If Google's been deleting it, over the last, you know, let's say it's the default that we think is 26 months. That means you won't have access to 26 months of this data, if you could access that through the API. The reality is if you don't have, do not automatically expire and we mentioned it before, but I'll put a link in the show notes. There's a couple of custom dimensions to put in the client ID, session ID, hit timestamp. If you don't have those in, you are not going to have access to this raw data on the free version of Universal Analytics. So I think there's this idea that, oh, we'll pull out all the raw data next year, but actually a lot of people don't realise that the data doesn't exist anymore or that they won't even be able to if they wanted to anyway.

[00:19:02] **Daniel:** So the reality for me would be that it doesn't matter because the kind of data you would want to export maybe are those standard report tables. Maybe you just want to pull out total sessions by day, total users by day. Even then, if you combine too many odd dimensions and metrics together, it's going to go and force a custom report and go back to the raw data, which means that in the defaulted ways, you know, you'd have access to 26 months of data. Which is a lot, you know, it is an awful lot, but if you haven't changed that retention setting, then you're not going to have access to your historical data, it's as simple as that. And I think this is why there's a lot of attention being focused on this right now, because a lot of people didn't realise it was there. It comes to a point in the future where, oh, I'm going to pull out my last 10 years of data and it turns out you can't, or you're limited to very, very basic data, which is maybe not even that useful at all.

[00:19:47] **Dara:** Well, that's exactly it, and I think that's something we've mentioned before, which is that it's very easy to feel the fear and the urgency of this and decide you need to export everything but the reality is a lot of that data is not going to be useful in any kind of meaningful comparison or benchmarking anyway. So chances are, if you haven't noticed the effect of your data retention settings, then it's probably not an issue, and you definitely don't need to worry about whether some of that data's going to be missing if you do a historical export. Okay, I think we've covered retention settings enough, this was a nice little ad-hoc, deep dive based on a real problem we were talking about. So switching gears, Dan, what have you been doing? I know you've been off for a week, so you should have loads of stuff to talk about in our little wind down section. So what fun things have you been doing outside of work?

[00:20:33] **Daniel:** I was off for a week, but I didn't go anywhere. I was just using the holiday and decompressing getting a bit of R&R, which was very nice but I was pretty lazy. Playing lots of video games, but yeah, enjoying myself, which is the key thing. However at the weekend just gone I did go to the theatre and saw a musical. So that's what I'm going to talk about now and that was a musical of the film, The Calendar Girls, I don't if you've ever seen that Dara? It's a film with Helen Miran and Julie Waters from the early 2000's and yeah, it's an interesting film, I highly recommend it. It's good fun to watch, but they turned it into a musical and I went to see that on Saturday. And so that was really nice, me and my wife went out for a nice meal beforehand and went to see a show afterwards. Really good fun, highly recommend it if you get a chance to see it then it's a really good adaptation of the film.

[00:21:16] **Dara:** Was that up in the London city?

[00:21:18] **Daniel:** It was not, it was in the Brighton town.

[00:21:21] **Dara:** How about you Dara, what you've been doing to wind down?

[00:21:23] **Dara:** Well I had, I had COVID last week for the second time. So that meant I watched a lot of TV while I was trying to recover. It was the only thing I could do, but I was on the mend by the weekend. In fact, I was in the all clear by the weekend testing negative again, and we went to see The Dandy Warhols play at the Camden roundhouse on Saturday night which was great. And then also, and I'm going to shoot myself in the foot and use this up, use two up in one week and probably have nothing interesting to talk about next time. But I also, we went to see the Elvis movie last night in the cinema as well, which was, which was great. The reviews are a bit mixed, but I don't always listen to the reviews and I thought it was, thought it was really good, really enjoyed it.

[00:22:03] **Daniel:** Ah, amazing. I've got a ticket to that next week so I'm looking forward to it. No spoilers, no spoilers.

[00:22:08] **Dara:** He's still alive.

[00:22:08] **Daniel:** Yeah okay, great, good.

[00:22:10] **Dara:** I'm looking forward to the sequel. Okay where can people find out more about you Dan?

[00:22:16] **Daniel:** On [LinkedIn](https://www.linkedin.com/in/danielperryreed/) and [danalytics.co.uk](https://danalytics.co.uk/?utm_medium=podcast&utm_source=the_measure_pod&utm_campaign=43).

[00:22:18] **Dara:** And you can find me on [LinkedIn](https://www.linkedin.com/in/darafitzgerald/) or daralytics. No, you can't that's not a, not a real website or if it is, I claim no responsibility for it. You can find on LinkedIn. That's it from us for this week, if you want to hear more from me and Dan talking about GA4 and analytics in general, you can check out our previous episodes in the archive at [measurelab.co.uk/podcast](https://www.measurelab.co.uk/podcast/?utm_medium=podcast&utm_source=transcript&utm_campaign=43). Or you can obviously just do it by using whatever app you're listening to this on right now.

[00:22:46] **Daniel:** And if you want to suggest a topic or a guest that we should be talking to on the show, then there's a [Google Form](https://docs.google.com/forms/d/e/1FAIpQLSeMwfzg9TdNKeZYRxHci_y0D9up4EyAL5zNMYZNE9P_82fz3Q/viewform) in the show notes, or you can just email us at [podcast@measurelab.co.uk](mailto:podcast@measurelab.co.uk).

[00:22:56] **Dara:** Our theme music is from [Confidential](https://open.spotify.com/artist/4erl5USs7RuiUU4QeiEQFO?si=xyGxD84PQHO78TKlvmZKLw), you can find a link to their music in our show notes. I've been Dara joined by Dan, so on behalf of both of us thanks for listening and see you next time.