Applying Analytical Patterns

Day 1 Lecture
Data Engineering Design Patterns at Meta - Growth Accounting

Transcript:

0:00
using analytical patterns is going to save you an unbelievable amount of time as a data engineer because let's face it
0:06
not all data engineering pipelines are built differently there's just a couple different patterns that you want to look
0:13
into and if you can see these higher level patterns then you'll know exactly what type of pipeline to implement in
0:19
this 2-hour course we're going to be covering two of the very important design patterns one is growth accounting
0:25
which is how Facebook tracks the inflows and outflows of active inactive users
0:31
this can also be used for any other state change transition tracking it is closely related to the cumulative table
0:37
Design Concepts in weeks one and weeks two that we used for dimensional and fact data modeling I'm really excited to
0:43
show you that and then we'll also be covering the Survivor analysis pattern which is looking at of all the users who
0:50
signed up today what percent of them are still active in 30 days 60 days 90 days
0:55
it's kind of like that retention number that is super important for the success andow growth of companies and if you can
1:01
learn how to apply these analytical patterns you will become a much more powerful data engineer uh I hope you
1:06
enjoy the course and if you want to learn how to apply these patterns in the cloud you should check out the data expert Academy in the link in the
1:13
description below and you can get 20% [Music]
1:19

off what are we talking about today repeatable analyses are your best friend and why are they your best friend one of
1:26
the big reasons they're your best friend is because they they allow you to think at a higher level instead of having to
1:33
think about like the group bu and the select and that that level of analysis
1:39
you can think one layer above that which is like kind of a more abstract layer
1:44
and then recognizing that uh the SQL will write itself once you can recognize
1:49
that AB that higher level abstraction and so if you have that higher level abstraction great that's great and
1:56
you'll be able to do some really good stuff with that and that's one of the big things as you kind of grow in your
2:03
career as a data professional you should be looking for these like higher level
2:08
abstractions that's why like you know when you start out as a data engineer
2:13
you work on like one Pipeline and then usually it grows to like a family of pipelines and then it's like a warehouse
2:20
of pipelines and then it's like the entire data warehouse and like you just keep building like a bigger and bigger
2:26
and bigger scope and so one of the things that you need to be able to do in order to get a bigger scope and get that
2:34
step up is you need to be thinking about things beyond the language that's written in and Beyond like the the ones
2:40
and zeros that it's written in because if you think about it uh from another layer right of why repeatable analyses
2:46
are your best friend is you can think of it as kind of like an
2:52
extension of SQL in some ways because on one side SQL or SQL I mean I'm not going
2:59
to call SQL so a SQL is going to be like uh already an abstraction because
3:06
SQL ends up just giving instructions to some sort of engine which usually then

3:11
processes it with like either usually like C++ or Java or some other uh kind
3:17
of more uh high performance language and then it will crunch the data down and then it will give you your data result
3:23
back and so that's also an abstraction so if that's an abstraction and working
3:30
with SQL like why do we have to be married to that layer why is that the layer that is the layer that matters the
3:35
most like is there such things as layers above and obviously there are layers above that you could think about that
3:41
but like a lot of these layers above are ones that I think are actually not good for your career uh for example um like a
3:52
low code or no code tooling where you uh just don't even write code at all and
3:57
you just drag and drop boxes around and that stuff is is useful but not generally speaking for technical people
4:03
that stuff is not as useful and we're going to go more into details around that and like how having overly abstract
4:10
stuff can actually be a detriment in some cases as well but anyways we're going to be talking a lot about some
4:17
common analytical patterns today uh one of the big ones we're going to talk about today is State change tracking so
4:24
State change tracking is essentially like one of the things I like about State change tracking tring is it's
4:30
actually really closely connected with uh SCD you can almost think of it as
4:37
like like the opposite of STD in some regards because you know STD has records
4:43
for every every different value of a dimension but imagine if you had a table
4:50
that instead of keeping all the values of the dimensions it keeps track of every time that Dimension changes and
4:57

you just have like the change log of like oh it changed on this day it changed on this day it changed on this day and that's kind of what state change
5:04
tracking is all about is all about like okay like where did they go like where did they go how did they change did they
5:10
come back kind of stuff like that uh we're also going to be talking about survivorship analysis uh which I needed
5:15
to put on this slide and uh the last thing we're going to be talking about is window-based analys analysis and we're
5:22
going to be doing all three of these on the lab today and they should be a pretty good time uh the the queries for
5:30
that stuff isn't they aren't too crazy and keeping in mind that some of these are really closely linked with the
5:37
concept of cumulation you have that cumulative table design that we talked about in week one and some of these are
5:44
actually based on top of cumulative table design and they don't work very well without cumulative table design so
5:51
keeping keeping that in mind as we kind of go through this presentation that what I'm trying to do is teach y'all the
5:56
kind of layering of all this stuff and hopefully this will give you some
6:01
powerful Tools in your tool belt for when you are building out your pipelines
6:07
so yeah let's let's dig into this stuff repeatable analyses right in the
6:13
last slide I was talking about how it reduces uh your cognitive load because you don't have to think about things in
6:19
SQL you can think about things that kind of like bigger picture layers right because imagine like if you're the CEO
6:25
of a company and you had to think about every single line of code like imagine well one of my beautiful examples I like
6:32
to give in this case is so at one point in my career I uh wrote a dashboard for Mark Zuckerberg and he liked my

6:39
dashboard a lot but like had I been like nazuk like you gotta you got to write
6:44
the sequel and he's not going to write the sequel because that's not his abstraction he's like n dude that's too much of a cognitive load I need I don't
6:50
need to care about the column names I don't need to care about any of that I just want the picture and so that's one
6:57
of the things that I hope hope that we as data professionals can get better at
7:04
is understanding the right layer of abstraction because a lot of the times I think as data
7:10
Engineers we are like we live breathe and die by SQL and if it's like if you
7:17
take our SQL away like some of us would feel like we are a Knight without a sword or we're a Knight without our
7:23
armor or where we are you know we aren't even bringing a weapon to the gunfight anymore if you take our SQL away and
7:30
so that's something I want to change and that's one of the goals of this presentation is for a couple of reasons
7:38
because one like in the future the abstraction is going to
7:43
change llms are going to change things SQL is going to be different or how we
7:49
create SQL is going to be different so really the more important thing is having these higher level patterns and
7:55
understanding how to apply these higher level patterns and then it's going to be a very in some ways it's gonna be a very
8:02
exciting time I know that in some regards some data Engineers have some fear about this where they're like oh
8:08
wow if a chat GPT can just write SQL like am I deprecated am I just going to
8:14
be laid off am I not going to be around anymore and I I don't believe that I believe that if you have a a firmer
8:21
grasp on these higher level analytical patterns then you can be that chat GPT
8:26

using engineer who can write pipelines incredibly quickly and your your value
8:31
to the company will be just astronomically higher and that's what that's why the second bullet here is all
8:38
about streamlining your impact because if if you had to think about like okay
8:44
if I needed to like go to the store and uh I want to make a cake and every
8:50
single time I want to make a cake I like go to the store get the bath or get like
8:55
the get the flour get the egg get the sugar get all like the the Raw materials and like you know Stir It Up and make
9:01
the cake like that's good and that kind of makes you a baker but that's not necessarily the the best way to go cuz
9:08
sometimes maybe you're strapped on time maybe the the stakeholders are like can you just pull this cake for me real
9:13
quick and you're like dude like cake cake making takes time and like if you you know if you ask for it to be pulled
9:19
real quick I'm going to burn it or whatever and that's where these patterns
9:25
can help you do things where instead of maybe making everything from scratch you
9:31
buy some of the pieces uh pre-made and then all you do is put the frosting on the top of it and then the rest of it
9:38
pre-made and ready to go so that stuff's pretty cool about that's what that's one of my favorite Parts about repeatable
9:43
analyses is once you recognize the higher level picture the pipeline writes
9:49
itself it really does like you have to be careful around the edges and like being careful around like cumulative
9:55
table design and like a lot of the gotas there for some of these patterns but if you got those down then I wouldn't worry
10:02
too much about it uh because you know the higher level pattern and what the pattern looks like and yeah you'll be
10:08

good from there so remember this repeatable analyses are your friend they
10:13
allow you to play with bigger Legos you don't have to play with like the little tiny two-piece Legos you can play with
10:19
like the bigger ones and then they can kind of Stack together and you can build a tower way faster so that's my main
10:26
goal here about this is uh recognize that repeatable analyses are very
10:32
powerful so I was when I was making this slide I was trying to figure out all the
10:37
higher high level analyses that I've ever done in my career and it's wild
10:43
because I really think that there's only three buckets here there there might be more uh uh like it if you are kind of
10:52
pushing the envelope and you're not in the data engineering realm and you're more in the data science realm and stuff
10:59
like that there might be some other uh analytical patterns that exist but for the most part I found there to be three
11:07
you have aggregation based patterns accumulation based patterns and window-based patterns and if you can
11:13
Master these three patterns you are going to be a master of analytics and like
11:19
95% plus of all the pipelines that I've written in my career in big Tech they
11:24
use these three patterns so if you can figure this stuff out that's going to be great like some regards you could say
11:31
that there's another one called like an enrichment pattern where you like do a join and bring in other columns but my
11:38
whole point about this uh lecture is that we are already at the master data
11:44
layer where we have all the columns that we need so enrichment does not take place in this realm because we already
11:51
have all the columns that we need so that's going to be the one that could potenti you could argue could also be in
11:57

this list so let's kind of go over each one and go over the depth of each one because there's there's a lot there's a
12:03
lot in each one of these buckets so I I really hope to uh illustrate to y'all
12:10
how to really build on these aggregation based patterns
12:16
so aggregation based patterns are probably the simplest patterns
12:22
so when you're building out a pipeline and you do an aggregation
12:29
there's uh a certain keyword a SQL keyword that be that should be screaming at you when you say aggregation there's
12:37
one keyword that should come up and that's Group by so you have group by uh aggregations are all about grouping by
12:45
uh different things and Counting Things based on different dimensions I remember
12:50
uh when I was working at Facebook they uh back in the day like when I was there
12:55
2016 one of the tenants of the company was was move fast and break things which
13:02
for me I loved I found that to be so amazing and then after like six months into working there they were like no
13:08
more no more move fast and break things we're done and then they tried to change it and then for like the software
13:15
Engineers they changed it to move fast and build things and then for the analytics people they moved it to they
13:21
changed it to move fast and count things and so like I don't know like I was
13:27
always sad about that but my main point with that whole story is counting is a a
13:34
very important part of analytics and it
13:39
it gets it doesn't get enough credit it honestly does not get enough credit because it's not sexy it's not fancy
13:48
it's just it's it's like I don't know it's like the meat and potatoes of analytics
13:53

like you know saying oh we have 7 million users in India and we have 5 million users in the US or d d like and
13:59
it's just like we have this stuff or like a line chart of growth and being
14:04
able to see like oh this is our line chart and you can do you can do really cool stuff with with line charts even as
14:10
well and you can be able to combine them and uh look at all sorts of different things and so when you're doing an
14:17
aggregation and an aggregation based analysis you're going to have some you're going to have some aggregation
14:23
functions like usually like sum or average or count usually it's those
14:30
that's usually it some average count that's that's about it like that's what you're going to be using for this like
14:37
maybe maybe uh more funky stuff like array a and other things but like but
14:43
we're already at the analytics phase so at this point you're going to mostly be summing averaging and Counting things
14:49
that's like uh in our last class I guess there was one more there was the percentile
14:55
percentile can be an interesting one but percentile and average are like the same
15:01
thing right percentile is just uh or average is just uh a specific case of
15:08
percentile where percent uh average is just 50th percentile so those two functions are technically the same thing
15:14
they just have like it's just like where in the distribution you pick and so I
15:19
don't know it's uh it's some pretty good stuff so uh but I no that's median my
15:24
bad median median and percentile are the same those are the same not average
15:30
average is different average is because average could be skewed made a mistake there but median is another one that you
15:36
could potentially use in these cases but one of one of the main things I'm trying
15:41

to say here with aggregation based analyses is you're going to group on things and you're going to count things
15:46
and you're usually going to bring in some Dimensions uh in this case you might have like country or gender or device or
15:56
I don't know all sorts of different dimensions and those Dimensions can be really powerful uh to see what's going
16:02
on and for example another thing I worked on at Facebook was this thing called the root cause analysis framework
16:09
where what it did was you could plug in any metric and then it would explain the
16:14
movement of that metric so say for example you had a week over week change
16:19
and the week over week change of the metric was like plus a million and then what it would do is it would give you
16:25
the the dimensional breakdown of that of like oh it's plus a million but it's actually like plus 1.5 million in the US
16:33
and minus 500k in India and it will give you the actual because just because uh
16:39
the total grand total of a metric is positive or like it moves in One Direction doesn't mean that all the
16:46
dimensional Cuts within it are moving in that same direction it's just that like in total they all add up to that
16:53
together so one of the things that you can do with root cause analysis is if
16:58
you see uh kind of a a certain shift in a metric that
17:05
you can start bringing in other dimensions to see like oh is it because of gender is it because of country is it
17:12
because of um device operating system is it because of height or whatever whatever or age or there's all sorts of
17:20
different dimensions that you could bring in to try to diagnose a shift in a metric and that is one of the things
17:28
that I think is one of the most important parts of being a data analyst is being able to do that is being able

17:33
to like say like okay we have a line chart here we see a dip and the reason
17:38
for that dip is x i an example I have is like I remember when I was early in my career at Facebook I uh was working on
17:46
growth and then there was a massive dip in growth and they did the root cause analysis and it was because Ethiopia
17:52
shut off the internet and so there was like no users in Ethiopia so and since
17:57
there's no users in Ethiopia that's a kind of a big problem and so that's kind of the idea behind root cause analysis
18:05
is you really are trying to do kind of the slice and dice and the storytelling uh that gives a better
18:12
picture besides like number go down sad it's not that it's number go down sad
18:18
and we have these people to blame or we have this product to blame and like it makes it so you can point a finger when
18:24
number go down and so that you can actually blame your sadness on something not just like oh no the numberers down
18:30
and I'm like I don't I'm freaking out man and so that's like one of the things that I think is very important uh when
18:36
doing grot cause analysis is it gives you a a lot firmer picture of what's going on with your metrics so yeah uh
18:43
let's let's let's let's go on to the next slide okay so remember what I was saying
18:49
before uh you have um it's the most common type of analysis like just
18:55
straight up it's the most common by far it's at least Le half maybe more of of
19:00
your analyses are going to be aggregation based analyses so and keeping in mind that those other two the
19:07
ACC cumulation and window-based analyses that I talked about uh in the earlier slides they can then be move they can
19:14
also have an aggregation based analysis on top of their own outputs so that's
19:20

where this is always just really really powerful and you know Group by always always have group by um so another big
19:28
thing um uh about this is you have uh trends
19:34
that we were talking about and then we have uh uh root cause analysis which I was talking about in the last slide
19:40
about like why why things are happening you can also have composition of like to know okay we have this many users in the
19:47
US we have this many users in China or whatever one of the things that I think is important about doing these analyses
19:54
is they should be at uh when you're doing these ation you shouldn't be going
20:00
back to the the fact data even though
20:05
there's probably uh a strong urge to go to the fact data and because with fact
20:12
data like the problem with it is fact data should be aggregated along the dimensional line usually that's going to
20:20
be user it might be like listing ID or device ID or like group ID or some other
20:28
entity mention but usually it's going to be user ID and you want to have things already aggregated up because like the
20:34
problem is is like if you're going to do fact data and you're going to like join in all these like spicy Dimensions as
20:41
well to bring in and then you're going to like bring all bring in all these spicy dimensions and then aggregate your
20:46
fact data like that's just going to be gnarly because those Dimensions won't be like one to one cuz if you're at the
20:54
daily grain where you have like one row per user per day per and and then one user one row per user per day per metric
21:02
then when you do the join it's the same number of rows right because you just do a join and it's just like one to one and
21:08

the matching is done and then uh the volume is better shuffle a lot better and you're going to get way better
21:15
performance if you have that kind of daily aggregation staging layer than if
21:20
you you know do your aggregations at the fact layer so always kind of think about that when you're doing this aggregation
21:27
based analyses is that you should have that other layer uh called like your daily data like your daily aggregate
21:33
data that is aggregated along like a user ID for a couple reasons one of the other big reasons is that if you have
21:40
that data that data is also very powerful for experimentation because then if you have okay all the data along
21:47
the user ID then you can put users in group a or group b or group uh C or
21:53
whatever test and control that you want and then it's very easy to then like add
21:58
Dimension into uh your metrics and that will show you oh uh people in group C
22:06
versus people in group b there's a there's a lift or a drop in the metric and that's where like you get it's like
22:13
a double whammy right you get your experimentation metrics essentially for free but you also get more efficient a
22:21
aggregation metrics as well for like slice and dice and moving things around so that's a big thing I needed to add a
22:27
chart here for composition but I think yall know what composition is I didn't need to put like an ugly ass pie chart
22:32
in this presentation for y'all to get what I'm trying to say so um anyways uh
22:39
that's kind of the idea behind aggregation based analyses so what about some gotas here
22:45
uh there's a couple gotas that I think are important is when you're doing
22:50
aggregation based analyses you and you're bringing in all these Dimensions
22:55
you want to be careful to not bring in too many dimensions

23:00
because uh so you know we have that daily data where it's like you have all my metrics but what if instead of that
23:08
what I did instead was uh I brought in these Dimensions I brought in dog name I
23:13
brought in city I brought in height I brought in uh profession I brought in
23:20
like follower count and then I just brought in like a couple different other dimensions that could be uh like but I
23:26
brought in all those at the same time and then the problem is is like once you get a certain number of dimensions in in
23:35
your aggregation based analysis you just get back to you just are back to the Daily data because you enough Dimensions
23:43
together will will go back to uniquely identifying a person essentially and so
23:50
that's where you want to be careful because when when you if you have like one person or a small group of people in
23:57
your like aggregation then I don't know does that matter it's like in those situations
24:03
where it's like oh this uh this grain of data uh had a 50% growth but it went
24:09
from four to six and it's like woo congratulations yay right but it's like
24:16
that's where when you're when you're doing these aggregation based analyses you really do want to be carefully
24:23
considering like the grain that you're working with that's where there's some grains that are beautiful like I really
24:30
like uh like like uh like gender is a great grain cuz it's like well it's going to cut it in two three four I know
24:36
like that many it's not very many uh and then uh also country is a good one country is an iffy one sometimes because
24:43
like some some countries are really tiny and there might be like you know for me I know that uh there's one person in
24:49
kyrgistan who follows me on substack and so uh if that person leaves then I'll

24:55
see a 100% drop in my metric and I'll be sad and but it's like you got to be
25:01
thinking about that as well that's why like when you are doing this stuff and you're looking at at at especially
25:07
percentage based metrics make sure you're looking at the actual counts as well because it's like oh it's 100% drop
25:13
but it's that goes from one to zero and it's like okay that's not a big deal then it's like one person so that's a
25:20
that's a big one um another big thing to look at here is when you're doing a long time
25:25
frame uh this is something I talked about a bit in week two when I was talking about like that long-term
25:31
analysis framework I built at Facebook is you don't want to have too many Cuts in your data when you looking at a long
25:37
time frame because time is also a dimension that's kind of like country
25:43
where it's like uh kind of high cardinality because if you are looking
25:50
at that kind of stuff then you want to be careful and that's where you can kind of if you want to do these analyses on a
25:56
longer time frame uh you can lower the cardinality of it and just don't look at it per day but maybe look at it per week
26:03
or per month or even per year where then you like aggregate all the data across
26:08
an entire year as opposed to like still holding on to the the daily grain
26:14
because if you hold on to the Daily grain and you're doing some sort of long-term analysis like you're just going to get a lot of rows a lot of rows
26:20
of data especially if you then choose other dimensions that also are high cardinality then you're going to get
26:25
really a lot of rows even like just cut country and day is going to give you a lot because if you think about it like
26:31

if you have 90 days and then you have country then that's like okay um that's going to be 90 time 200 that's not too
26:39
bad 1,800 rows for 90 days right but then if you add any other like other high higher cardinality Dimensions good
26:45
luck right because then you have 1,800 times whatever other dimensions you also want to cut by so that's what I'm saying
26:52
is that like the combination stuff really does make a big difference so
26:57
yeah aggregation based analysis pretty good uh pretty straightforward I hope uh that stuff is important and uh my whole
27:05
my whole point here is just trying to give you a perspective at a kind of a higher level of not like oh I need to
27:10
solve this specific group by query because I'm sure most of y'all in this call have done this exact set of stuff
27:18
before but haven't maybe thought about things at a more abstract layer like what I'm talking about
27:24
now okay so now we're going to switch gear here uh to cumulation based patterns for
27:31
cumulation based patterns we're going to talk about two big ones uh we're going to talk about State transition tracking
27:37
and then we're also going toh talk about retention and uh J curves it's called
27:42
retention it's called J curves it's also called survivorship analysis there's a bunch of different things where I'm
27:48
going to have a table that shows all the different like things that you can do with that one uh and then State
27:53
transition tracking has some other names as well so yeah let's kind of dive into this so these patterns are going to all
27:59
be based on that cumulative table design that we worked on in week one and so if
28:05
you have a firm grasp of that fundamental these analytical patterns can be built on top and uh I think
28:12
you'll be very happy to see how this works okay so for cumulation based
28:17

patterns uh one of the big things that is different about them versus uh aggregation based patterns is
28:26
that they care a lot about the relationship between today and yesterday
28:32
and the shift of State between today and yesterday and so those things can make a
28:38
big difference where it's like okay are you active today versus yesterday or not and kind of how you shift and change and
28:45
do that kind of like slowly Dimension kind of uh as you walk through your life
28:51
and so that can be uh like that's like probably the biggest thing that you are going to be looking at when like or the
28:57
biggest difference between like a simple aggregation based uh pattern versus a cumulation based pattern and remember
29:04
full outer join here your best friend it's going to be so good because the big reason why you need full outer join for
29:11
these patterns is because you need to keep track of when there isn't data
29:17
that's the other big thing that's different between this pattern and the aggregation based pattern is that no
29:23
data is data and like it's like that's the like that's of the things that like
29:29
aggregation based patterns don't care about like if there's no data you just ignore it and it doesn't matter but in
29:34
cumulation based patterns no data is data because the fact that they didn't do something means that we want to keep
29:42
track of that the non-existence of data is something else that we want to keep track of so that's a big thing as we
29:48
kind of go through this like I think that will make more sense as this presentation goes on uh so I was talking
29:56
in the last slide about the survival analysis and state change tracking so uh those are kind of the two big uh
30:02

patterns that we're going to be covering in the lab today we're going to talk about growth accounting real quick this
30:07
is a a special version of the state transition tracking and this is where uh
30:15
the cumulation part of it will make more sense so you have in this case uh
30:22
there's five states a user can be in so uh the first state is new which means
30:29
that yesterday they didn't exist and then now today they are active which they have to be because they signed up
30:35
today so uh that's uh going to be the first state then you have um retained
30:42
means that they were active yesterday they're active today pretty straightforward they just stuck around
30:49
churn is active yesterday inactive today so they like left uh resurrected is like
30:55
the opposite of churn where they were inactive yesterday and active today and then stale are the people who like they
31:02
didn't come back they were not active yesterday and they're not active today so they churn out or or they are stale
31:07
because they don't they don't care in some uh in in some patterns you have you have
31:16
a you have a sixth state of like deleted or deactivated for users who are who are
31:25
like active or inactive today and then they um they don't exist or they were
31:30
active or inactive yesterday and then they don't exist today because they like disappeared because like you have to
31:35
filter them out because sometimes in these growth accounting patterns you have to there's like policies that make
31:40
it so you can't just hold on to everyone's data all the time because like that's like a privacy concern so in
31:46
those cases like sometimes you can have a sixth State here which is deleted um
31:52
Facebook just anonymized everything though so that we could just hold on to everything forever so and if you hold on to everything forever than there

31:58
actually isn't a deleted State you just are stale you just stay stale forever
32:03
and so yeah that is I I've noticed that that's one of the things that I've noticed that has been a difference
32:08
between this pattern like or that like how this pattern is different uh depending on how you apply it
32:15
so one of the things I want to talk about here though is this is this
32:21
pattern is actually incredibly incredibly powerful so let's kind of go over like a chart with this though cuz I
32:28
think that will help make this pop so this is kind of like the idea behind
32:35
how these charts work right so um so this is like ma monthly active users
32:40
growth accounting and in this case you'll see you have churned for all the people who are leaving those big red
32:47
bars and then you have new right these are the new users and then you have the resurrected users and so in this
32:56
case growth you can think of growth as new plus
33:01
resurrected minus churn and that will give you the growth rate of your
33:07
business or the growth rate of how many people are coming back how many people are are uh or how many people how many
33:14
incremental people you're getting because it's like the people who are coming in and the people who are leaving
33:19
so you have like the the incoming and the out so one of the big things that I want to emphasize here is
33:28
this is um not just specific to growth you
33:35
can do this in all sorts of different areas as well because you could think of
33:41
like churn resurrected new these are just different states that uh an entity
33:47
can be labeled in so I want to talk about just a couple different other times so I've done growth accounting at

33:52
Facebook I've used this exact pattern at Facebook uh to track like notifications growth and stuff like that to see how
33:59
users are using different notifications and interacting with different notifications but like and kind of the
34:05
kind of the the poster child version of this pattern but I've also used this pattern for uh more abstract or or or
34:12
more different things that like for example also at Facebook when I was working there I worked on tracking fake
34:18
accounts so in that case you have like a new fake account which was an account that had never been labeled fake before
34:25
and then they were labeled fake for the first time and then you have a a resurrected I call instead of
34:30
resurrected I called it a reclassified fake account and that's where you have a fake account that like was labeled fake
34:37
and then they were unlabeled fake because they passed the challenge where they like uploaded an ID or something
34:42
like that they actually were like see I'm I'm a real person and then then they were classified fake again so that's
34:48
like fake reclassified and then you also have like the churn state which is Declassified
34:54
and in this case churn is like I mean that could be a good thing actually in Facebook that means that there's like
35:00
fake accounts that are leaving right and so in that case you can see for fake accounts you have the exact same set of
35:07
States as you do for growth but it's just fake not fake and like like like
35:14
the history of like were they ever classified that way ever so that's
35:20
another great example of like a use case for where you can use this state
35:26
transition tracking so I want to go over some more that I use this this this analytical pattern so at Netflix uh I
35:33

had another kind of classification algorithm because you could think of like the fake accounts right that's a that's an ml algorithm it's a classifier
35:41
right where it labels things fake not fake done right and then at Netflix there was another machine learning uh
35:47
algorithm I worked with that was labeling uh applications in the in the
35:53
the Netflix's microservice architecture it labeled them as RIS risky or not risky or like too risky and it was like
36:01
a bunch of different states like that where it was like this is this is like critically risky versus not critically
36:06
risky and then uh we wanted to track the flow of that as well because one of the
36:11
things that this these types of patterns gives you is it gives you very good
36:16
monitoring you can monitor the health of your machine learning models because
36:22
when you are tracking stuff this closely if like the fake account model at
36:29
Facebook like say they push a change and then it says like oh look at that like all these new accounts are getting recl
36:35
all these accounts are getting like reclassified now or there's new accounts or all these accounts are getting Declassified and that's not what we
36:41
expected like this is a very very powerful uh way to model the health of
36:47
your machine learning models so and I had one more I had one more time that I
36:52
use this pattern this pattern has been very very very important in my career is at
36:58
Airbnb uh I worked on another model that um was for hosts because I don't know if
37:05
y'all have ever been on Airbnb and ever had the experience where a host cancels on you and like you fly to Europe and
37:12
then the host is like sorry bro uh we can't host you actually and they cancel last minute and that's actually one of
37:18
the most painful things on Airbnb to happen and like arbb is really trying to
37:23

prevent that so you know you can have ml that classifies hosts B based on the probability that they're going to do
37:29
that and so that was another one that we did the same tracking of like it was
37:34
like essentially was labeling hosts as risky or not risky and then kind of
37:40
seeing the health of those and trying to see like okay what are we doing and how are we because this is good for two
37:47
things because on one level it's like you can look at these charts to track the health of the model you can also
37:53
look at these charts to track the effectiveness of whatever your doing like for example like on Airbnb it's
38:00
like are our higher level goals of like educating hosts and teaching hosts are those things having an impact on host
38:07
Behavior so that we have fewer risky hosts and so you have that kind of like higher level strategic thing that you
38:14
can measure as well with this pattern so I know I just hammered home that this
38:20
pattern is really important but like I've seen this pattern be very effective like everywhere I've worked in big Tech
38:26
and so I and I generally think that this pattern is a very good pattern for ML
38:32
Ops and like if you especially if you have a machine learning model that is a classifier so if you have a classifier
38:38
machine learning model that's great and that would that this is where you can use this a lot and that is where yeah so
38:45
that's growth accounting or state trans transition accounting I love this chart and we're
38:51
going to talk about survivorship analysis and bias so this chart here is
38:58
in World War II there was um fighter pilots would go and fight and then they
39:04
would uh they would fight fight fight and then they would fly home and then they would look at the the planes and
39:10

then they're like wow uh we need to uh fix the plane and
39:16
initially people were like okay and you see this chart all the red dots are the
39:21
damage are like the where it got shot and like the plane experienced damage
39:27
and people were like wow we need to uh we need to bolster up we need to make it
39:32
so that the areas with all the the shots are stronger so that they aren't shot
39:38
but then a really smart man was like but wait a minute but those are the planes
39:43
that survived so actually the areas on the plane that need to be bolstered are the
39:50
places with no bullet holes and that will increase our chances of survival because that
39:56
will bolster the areas that can't take a hit and so in this case you see the
40:01
areas on the plane here like here like kind of that midsection of the wing is where there's like no shots that's
40:08
really where they need a bolster because if a plane takes uh damage there they don't come back and so there's actually
40:15
it's it's interesting how survivorship bias can impact you but in this case it
40:20
makes sense it makes a lot of sense that that's where you would want to go if you wanted to make your planes more effective and like that's one of the
40:28
things that I found interesting about a lot of different things like in my career and in other places right where
40:33
it's like I think like in content creation that's a big one where it's like people like they like content
40:39
creators like they will go on their journey and then like they get big and then they uh like they're like wow just
40:45
just do what I did and it'll you'll be successful and then a lot of people I think have a lot of false hope sometimes
40:50
because they see the only people they see are the stars they only see the people who won cuz you don't really
40:56

necessarily see the people who tried and failed because they are kind of more invisible and so you want to be careful
41:02
survivorship bias can really have a gnarly impact on your uh view of the
41:07
world and uh view on like what's considered fair or what's considered normal or all those kind of
41:15
things and that was a massive tangent for this presentation but my whole point with this is survivorship and like how
41:24
long things survive is another important measure measurement that we need to have
41:29
in our analytical patterns so let's talk a little bit deeper about this
41:35
okay so if you think about retention retention in some respects is
41:43
just Survivor surv surviving so you can see how there's essentially three ways
41:50
that things can go uh and you see at the top here how everything starts at 100%
41:56
because because there's the the fundamental component of this analysis is what's considered a
42:04
cohort or a reference date so it's a date that like everyone is on the same
42:10
page a lot of times for growth that's like the signup date for a user and then
42:15
as time progresses the state will change because some users will stick around and some users are going to leave and so
42:23
generally speaking if you have an app that has that Gray Line your app is kind of doomed because it's
42:29
like as you get new users like over time they're just going to go away forever and you don't have that kind of like
42:35
stickiness that an app needs to like survive you need that like crowd of people that just keeps coming back and keeps coming back and keeps coming back
42:42
so that's where like if you have either a green line or an orange line those are going to be the cases where you have
42:50
found a successful app because there's a certain percentage of users that are sticking around for the long term
42:55

especially that green line where like people come back and then like there there's a bump in retention over time
43:00
where people actually uh come back and they are more engaged than they were on
43:06
the day they signed up and that can be a very powerful thing to see um so J
43:11
curves and the reason why they're called J curves uh is if you kind of like imagine if you flipped your laptop on its side you would see there's like a
43:18
little J that you can kind of see there uh that's why it's called J curves so what are kind of some applications of J
43:24
curves maybe Beyond growth right there's it's obvious ly like I don't just like to talk about growth analytics because
43:29
there's so many other areas where you can do jcurve analysis as well
43:35
so essentially J curve analysis has three points you have the curve you have
43:41
the state check and then you have the reference date so in uh you could think of uh for users
43:49
you have a user who um users who stay on the app and then you have the state
43:55
check is are they active or not and then the reference date is when they signed up you can also think of like cancer
44:01
patients uh uh the the state check here is that they are still alive or they're not dead and then uh in this case the
44:09
reference date is going to be their diagnosis date that's where you know like how they have like the those
44:15
different metrics around like fiveyear survival and 10e survival all those
44:20
things because I remember like uh like like one of my relatives got diagnosed with like pancreatic cancer and that
44:25
cancer is like really terrible like it's like the 5year survival rate is like like 3% or 4% or something like that
44:32
it's like essentially a death sentence uh because the survivorship analysis is so it has that really it's like one of
44:38

those gray lines right where it's like you just experience that massive massive Decline and like nobody gets out alive

44:44

because it doesn't flatten out for anybody and so that could be the case other cases you have um you can think of

44:51

like the smokers Who Remain smoke free after quitting so you can kind of have maybe the opposite of more of a more

44:57

positive way of of going about it where you have uh that's the state check are they are they not smoking and then uh my

45:05

favorite obviously is uh you know in my in the when I launched this boot camp uh I I said if if y'all attend all the

45:12

sessions and you do all the homework then I will write you a recommendation and there was a lot of y'all who were like really cocky about that and you're

45:18

like I'm definitely going to take Zack up on that and now that you've listened to me talk for like 20 hours you're like

45:25

okay this is a lot more than I thought was going to be um but then obviously there's a little J curve there as well

45:31

but the thing that's interesting for me is I'm already starting to see that flatten and I I already have a pretty

45:37

good idea of how many recommendations I'm going to need to write but that's kind of a kind of the perspective that I

45:43

would uh think about like these survival analyses is like a state it's it's like

45:49

a state that is prolonged throughout time that is checked versus a reference

45:55

date it's the good way to think about it okay

46:01

so let's we have one more section here which is window-based analyses which

46:06

won't take quite as long because there's not as much in here as there is in other

46:11

cases but so for window-based analyses you really uh are going to be the the

46:17

first one here I I love calling is like it's it's the dods and moms right it's like I always think of like Mom and Dad

46:24
are the freaking that's the first I have day over day week over week month over month and year over year uh where you're
46:30
looking at I I like to think of that as more of like a derivative right it's like a rate of change over a period of
46:37
time essentially and that that's where using window functions is great it's so
46:43
great and then you have literally the opposite as well and I love I love how like window functions do both because
46:49
then you have rolling sum which is like the integral and it's like uh the the the cumulation over a certain window and
46:57
so you get that as well and then obviously rank as well rank is interesting but rank doesn't need to be
47:03
solved with window functions you can also just use like a plain Jane like order by as well but it depends on how
47:10
complicated of a rank you're trying to do so how does uh window-based stuff
47:15
work so one of the key words here is rolling right you're going to be looking for Rolling or week over week day over
47:22
day something like that and then but for Rolling you'll see that there's a specific uh syntax it's literally this
47:30
syntax every single time and then you just change n here so you change usually
47:36
sort here is going to be just you sort by date you Partition by like I don't know user ID or you Partition by uh
47:43
whatever um the dimensional cut maybe you Partition by country or you Partition by whatever your uh whatever
47:51
you're trying to cut by right whatever the the grain is and then uh the N here
47:56
here is going to give you the number of rolling days and it's that easy the straight up is
48:04
that easy that's that's what you do and uh like that's how you get your rolling rolling sums and the cool thing about
48:11
these is that like essentially they do the opposite of each other so if you do

48:17
the uh derivatives like the day over day week over week month over month year over year your line chart is going to
48:24
get spikier it's going to get spikier and it's going to be more uh sensitive to change especially day over day day
48:31
over day is going to be really uh noisy and then uh if you do month over month
48:36
it becomes less noisy and then year over-ear a little bit less noisy but year-over-year is still going to be more
48:42
noisy than just the regular chart like the the count chart because that's the
48:48
nature of it it's a derivative right it's the rate of change of the thing because now you're relying on two data points you're relying on what it was
48:55
today and what it was a year ago and one of the things that I found always so interesting when I worked at Airbnb is I
49:01
I worked at Airbnb in a very weird time because normally in big Tech they love year-over-year year-over-year Is Just it
49:08
gives the right uh it gives the right like acceleration view for a lot of
49:13
different metrics but I worked at Airbnb in 2021 and uh in 2020 Airbnb had a hard
49:21
time like because of the pandemic so we actually for a while Airbnb actually changed a lot of their metrics where
49:28
instead of looking at year over year they were like we're going to do year over two year so they compared uh 2021
49:33
to 2019 instead of comparing it to 2020 because if you compare to 2020 it was just like we're doing great we're we're
49:41
up 600% or like whatever right because it's like everything was shut down in 2020s so sometimes these me these these
49:49
metrics can be gamed like that right and that's that's why those derivative metrics can be so uh loud and so sens I
49:57
because if there was a dip or spike a year ago or a dip or Spike today then
50:03

you're going to see a Dipper Spike uh today right for year-over-year so it's
50:09
essentially like you increase the volatility of the chart if you are doing day over day week over week month over
50:15
month whatever right but you can also decrease the volatility of the chart with using this rolling sum and rolling
50:22
averages you you can when you integrate right and that's how you get a chart that kind of looks like this so I don't
50:28
know if y'all have ever uh traded stocks before I I used to do that a lot um and
50:34
so you'll see uh this uh the line here is going to be the ups and downs of the
50:39
stock price that is normal and then you see this blue line is actually the the
50:45
50-day moving average of the stock so and you'll see the 50-day moving average
50:50
is really uh different and it's very smooth and it doesn't like the it's more of like a trend line than uh what the
50:58
actual value of the stock is at that moment in time and the thing is is like you can have an even slower trend line
51:05
right because you can have a a 200 day moving average and a lot of times in
51:10
stocks that's what they look at they look at the 50 day and the 200 day and when the 50-day crosses above the 200
51:17
day that's called like a golden cross and that's like a a signal to buy and then when the 50-day crosses below the
51:24
200 day that's uh considered death Cross or like a signal to sell and so you can
51:30
have like uh these charts that like or these trend lines that are based on the
51:35
actual data but they kind of remove some of the the volatility or some of the you
51:42
know the day-to-day fluctuations of the charts and that can be very good like
51:47
that can also be a very good thing because that can help uh reduce the noise and make the picture a lot clearer
51:53

of like what's actually going on with this this this trend so one of the things about
52:00
window-based analyses that I want to be uh just just a quick side note is make
52:07
sure to partition on something like uh if you're using big data it you should
52:12
probably partition on something uh because if you don't like your window is going to be so huge and then uh that's
52:19
going to cause o to happen I've noticed this where like you're going to need a partition on some sort of dimension for
52:25
the most part like unless you're just trying to do like the rolling average of like all users or something like that and it's like that can be like you're
52:32
gonna need to like either crank up your spark memory or something like that like just so just a quick little side note
52:39
about uh these window-based things in the Big Data environment that you want to be aware of congrats on getting to