Data Visualization and Impact

Day 2 Lecture

Transcript:

for the day two lecture let's talk about a couple uh couple things that I've noticed uh bad practices that people

Bad Practices

48:27
have done um when they are building their dashboards so a big one is don't
48:34
uh don't do joins on the Fly Like if you have two data sets and like you're
48:39
trying to query something and you're like give me this data and then do a join and then dump it out you can do
48:47
that like if the join is Tiny like if it's like one side is super tiny and it's not going to be very uh not a big
48:54
cost but if the join is like any if both sides of this joint are like
49:00
substantial at all then don't do joints just don't do it like just pre-aggregate
49:05
denormalize and pre-aggregate like when you're in the dashboard you shouldn't be
49:11
thinking about master data or scalability or freaking any of that
49:16
because you're at the final layer because when you're in the dashboard like there's nothing Downstream of the
49:22
dashboard besides the human and so in if you have someone looking at it like then
49:29
it should be as crunched down and refined as it can be cuz there's not going to be anything that's going to be like reading from your dashboard so
49:37
don't use don't join on the Fly it's terrible um also like if you can do
49:42
pre-aggregation pre-aggregate your data sets I think that's another really powerful thing that can uh make your
49:49
dashboards really really really performant um we're going to not like there's actually a trade-off here for
49:54
pre-aggregated data sets because on one side you have um when you preaggregate

50:00
you lose like a certain layer of detail that you might have if you have all the
50:05
dimensions of all the individual rows but on the flip side it just loads
50:11
instantly so like if you like especially I remember at Facebook when I was building out like I built out a lot of
50:17
very fancy visualizations at Facebook I built out the family um family of apps
50:23
visualization at Facebook which is I was the very first Eng ER at Facebook to integrate all the metrics from Facebook
50:29
WhatsApp Instagram and messenger and we had like one visualization that showed all of that data and like if you think
50:36
about it if I didn't pre-aggregate then the number of rows that each day of data there would have
50:44
would be like four or five billion for each user and it's like the dashboard's
50:50
not going to load fam it's not going to load because it's like that's for one day like think about like oh we want to
50:56
show show a 30-day line chart okay now instead of 5 billion rows you have 150 billion rows right and it's like good
51:03
luck right so in those cases you want to aggregate down and then you can aggregate down to like a certain grain
51:09
and if y'all remember uh we talked about in the advanced SQL section of this boot camp we talked about grouping sets
51:16
grouping sets is very very very powerful in making your dashboards a lot more performant because then you can pick the
51:24
dimensional grain that you want you can have multiple dimensional grains that you can look at and you can build a lot
51:30
faster dashboards that way so definitely things to think about um also uh another
51:37
thing that I've noticed is some sometimes uh I saw this at Facebook and at other places is they people like to
51:45
use like like stupid stores for their data like they like to query S3 directly

51:53
which S3 is like kind of cold like you know how like they say Apache Iceberg sits on S3 and iceberg is I don't know I
52:01
don't know if you know about icebergs are not hot icebergs are pretty cold like you know like if you like you know
52:07
watch the Titanic or whatever and like Titanic crashes into the iceberg that was in the Arctic kind of cold not fast
52:13
not hot so you want to do not using S3 you almost never want to build your dashboards on top of S3 um you want to
52:20
move the data into a lower latency store uh Druid being the best example um and
52:25
obviously now since jender was here I'm going to have to make a shout out to penino that sounds like another really
52:31
cool option that you can use as well so make sure that you do that cuz like if you do if you do both if you like use
52:39
Druid and you pre-aggregate and you don't use joins you do all three of
52:44
these things your dashboards will load instantly always regardless of the scale
52:49
regardless of the complexity they will always load instantly and they and even if people change filters they change
52:56
stuff like that you will have a very fast and performant dashboard that people will love using so remember to
53:03
think about the end user when you are building out your dashboards because if
53:09
your dashboard is slow or kind of hard to use then uh people are just going to not use it I have a statistic for you
53:16
you know that 85% of dashboards in um in uh businesses are used one time or less
53:24
and then never used again so that's like a terrible terrible
53:30
number just letting you all know that's a terrible number like and so um and the big part of that is because they don't think about these things they don't
53:36

think about investing in quality or completeness and then but then obviously the dashboards that are used can be used

53:42

a lot so you want to uh understand the different ways that things can work here but these are your best practices so uh

53:49

remember remember those uh these things are not as important to remember as don't use pie charts but they're up

53:54

there okay how to build dashboards that make sense um so one of the things that I

Who is your customer

54:01

would do uh when I was building out my dashboards at Facebook was I was thinking about okay who is this

54:07

dashboard for and um if if your dashboard is for

54:12

executives it shouldn't have any interaction for the most part like it should just be like your charts your

54:19

lines and you have one story to tell like essentially if you're giving a dashboard to an executive

54:26

the dashboard and a screenshot of the dashboard are the same thing and that

54:31

like there's not like you don't have like as much like interactivity analysts on the other hand want almost the

54:38

opposite they want tons of drop- down filters they want to be able to slice and dice and kind of like really hunt

54:45

with the data sets so this is where um for analysts like if we go back to the performant dashboard uh kind of um best

54:53

practices uh you might end up not preag ating for analysts because you might be able they might be able to have more

54:59

complex filter conditions if they are um not preaggregated and you can have like

55:05

uh they can kind of do more root cause analysis if it's not pre-aggregated and

55:12

for them they understand if the dashboard is a little bit slower because it's doing something more complicated so

55:19

just a thing to think about the big thing to remember here is who is your customer uh if you build uh it kind of
55:25
works both directions here it's like if you build an executive dashboard for an analyst they're going to be like well
55:33
thanks for the one number but then it's like the other way around it's like if you build an an exploratory dashboard
55:39
for an exec it's like thanks for like all the filters that I'm never going to use and so they both uh kind of cut both
55:47
ways so definitely um that's going to be a big thing to think about so knowing
55:53
who your customer is really does uh determine the design that you would be thinking about okay what are the types of

Types of questions

56:00
questions that could be asked so let's talk about a couple different types here these are going to be uh the main types
56:06
of questions that you are going to surface up when you write your dashboards and do your visualizations so
56:13
uh first one here is going to be uh Topline questions like how many users do we have how much money do we make right
56:22
uh how much time are people spending on the app like stuff like that and you have Trend questions like how
56:29
many users did we have this year versus last year right and like and you could
56:34
think of like Trend questions they almost always have that uh they have some sort of uh time component to them
56:40
that's going to be the one thing that's very important with them whereas Topline questions have essentially no Dimension
56:46
right or they might have a filter but that's about it they might might be like how many users do we have total in India
56:52
right or something like that that would be like a kind of filtered top of line but that's kind of like uh that's where
56:58

it's interesting between top of line and composition because composition is going to be like okay what percent of our
57:03
users are Android versus iPhone or like uh different things like that where you
57:08
have like what percentage of our users is male versus women and uh different things like that so those are good we're
57:14
going to go a little bit deeper into each one of these types of questions because they're they're all very kind of important um

What numbers matter

57:21
but what numbers matter here so these are going to be a mix of top of line composition in Trend so like you know
57:27
you have total Aggregates right total Aggregates is just like the total number right I love the the number I love there
57:34
is the total aggregate I like to talk about is um if you uh just do count star
57:39
on from number of humans or from All Humans on Earth with no Dimensions right
57:45
you get like like a 100 billion there like there's been like a 100 billion humans on this planet that's a lot of
57:51
humans um and uh then you have time based Aggregates right so you could say
57:56
like um in that casee you could say like okay you could do count star uh but then
58:02
in that case from humans but you Group by like this year and then instead of
58:07
getting a 100 billion you get like 8 billion maybe a little bit more because a lot of people die this year as well so
58:13
like it might be 8.1 8.2 or something like that but like uh like you get your time based Aggregates that way and then
58:20
uh you get time and entity based Aggregates where in this case you could say like okay how many alive humans do
58:25
we have have so then you do just get that 8 billion number uh or it's like how many alive humans do we have this
58:31

year that's like a good thing to think about is like that's another aggregate kind of metric that you can have um you
58:38
have derivative metrics so this is a good one where it's like uh how many this is like year-over-year how many
58:44
humans are there like so you know in the US like uh what we we're like plus a million or plus 2 million or something
58:51
like that like that's how many more humans are on in the US this year than compared to last year so you get that
58:57
like derivative metrics are very powerful they are um when I was working at Facebook and I built a dashboards for
59:04
executives they actually didn't care about the total numbers at all they
59:10
actually asked me so the original dashboard I had for them was the total numbers like you know Facebook has two
59:16
billion WhatsApp has 1 billion messenger has 1.2 billion or whatever it's like just those up and to the right charts
59:22
and they didn't care about that at all they only cared about year over-year like like was the growth accelerating
59:28
decelerating and they they cared a lot more about the derivatives because so derivative metrics are interesting
59:34
because they are a lot more sensitive uh to change than um just
59:41
normal Aggregates are so like if you look at a chart and you look at the derivative it's going to be a lot more
59:48
like volatile than just the the normal aggregate is so and that could be very
59:53
powerful in charting because it can help you pick out Trends earlier uh and that
59:59
could be a big thing I don't recommend day overday derivative metrics because day overday is like you have a lot of
1:00:05
weirdness especially with like Friday and Saturday and then you also have Monday and Sunday and like because
1:00:12

there's weekend versus weekday and there's usually like a a seasonality to the week and so like your day overday

1:00:19

metrics are usually pretty uh volatile and insane and don't really have a pattern unless the unless the trend is

1:00:26

very very strong and then uh you have other types of metrics dimensional mix

1:00:32

like for example my in my my newsletter I have 33% of my subscribers are in the

1:00:39

US 25% are in India and then the like the next biggest country is like 4% and

1:00:46

it's Brazil I know Bruno's in this call so shout out to Brazil uh and then we

1:00:51

have um Android versus iPhone uh that's going to be another big one uh that

1:00:57

people like to battle about and there's all sorts of other dimensional mixes that you could think about uh and then

1:01:03

we also talked about retention and survivorship that's something that uh we talked about in the analytical stuff as

1:01:09

well which is its own type of metric which is like it's it's kind of like a dimensional mix though it's like the

1:01:15

percent of users who are still this Dimension after this amount of time so

1:01:22

um one of the things I want to go over real quick about dimensional mix that can be interesting is sometimes these

1:01:29

numbers can change but like the dimensional mix can change but then the

1:01:35

total aggregate stays the same like for example uh when the Ukraine war happened

1:01:43

uh I'm sure the number of Facebook users in Ukraine uh dropped a lot but they

1:01:50

didn't uh they didn't drop overall because they just left right they moved

1:01:56

they moved to like Poland or they moved to like Russia or they moved to other places in Europe and they kind of fled

1:02:02

Ukraine and so like that's what's called mix shift which can uh which can happen

1:02:08

where it's like if you have a a metric that's tied to a dimension then uh

1:02:13
sometimes that that Dimension specific metric can drop but that drop actually
1:02:20
doesn't have any impact on the business because the value just changed it's like
1:02:25
because you're tying your metric to a slowly changing Dimension and if you tie your metric to a slowly changing
1:02:31
Dimension and then that Dimension changes H that user is still there it's
1:02:37
just that they are not they don't have that same dimensional value anymore so you want to be that's one of the things
1:02:43
you want to be careful about when you're defining dimensional mixes because that
1:02:48
can happen more often than you would think like it happens all the time so um
1:02:54
that's a good thing to think about like mix shift so these are the most common kind of uh visualizations and charts
1:03:01
that you would see in dashboards so um yeah let's go to the next slide why do
1:03:07
these numbers matter total Aggregates we're going to talk about each one of these uh in more detail here so total Aggregates is almost always reported to
1:03:15
Wall Street uh because it's like uh you know Airbnb always reports number of
1:03:20
bookings bookings is the number that Airbnb reports Facebook reports um
1:03:26
active users but Facebook changed right so here's a great example of how Facebook changed their reporting to make
1:03:32
Wall Street um happier in some regard is Facebook no longer reports on the number
1:03:39
of users on the Facebook app they only the only number they give to Wall Street
1:03:45
is how many users they have on the family of apps so on Facebook Instagram
1:03:50
WhatsApp and messenger they don't give the app bya breakdown they only give the
1:03:56
total aggregate to investors one of the reasons for that is because the the Blue app like the original Facebook app like
1:04:03

wasn't doing so well all right and uh there was a couple times when it didn't do so well and investors fled Facebook
1:04:10
right and so this is goes back to um when we go back to like what I was talking about in this previous slide
1:04:15
about dimensional mix is uh you have dimensional mix and mix shift and a lot
1:04:22
of times these users are still actually on a Facebook Prof product they just moved from the Facebook Blue app and now
1:04:29
they're on Instagram or they moved from the Facebook Blue app and now they're on WhatsApp right and they aren't they just
1:04:35
aren't on the Blue app anymore but meta and Facebook Facebook meta whatever they they still actually have that user they
1:04:42
just don't have them on that app specifically and they're not tied to that Dimension anymore so that's a great
1:04:48
example of where when you're reporting these total Aggregates a lot of the time
1:04:54
uh you don't want to to report uh Dimension specifics because those
1:04:59
Dimensions can kind of change and uh these total Aggregates are very important it's like how much money are
1:05:05
they making and then uh and all that kind of stuff right and these kind of just give a current current state like
1:05:12
obviously another one is like Revenue how much revenue they made in the last quarter that's another very common
1:05:18
metric or aggregate that gets reported so think about total Aggregates good good numbers to think about you have
1:05:24
time based agre tet uh these are going to be um a little bit different right
1:05:30
where you have you you can catch Trends earlier this way uh you know a bad quarter is the potential signal of a bad
1:05:37
year you can see that happen I mean I don't know if yall saw Facebook stock over the last year it like went from it
1:05:42

went from 300 to 90 back to 300 some crazy stuff and these charts uh you know
1:05:47
this all about growth and Trends and like where is the stock going that's what these time based Aggregates are going to be
1:05:54
um so so time and entity based Aggregates these are not usually reported to Wall Street uh but they are
1:06:00
often plugged into ab AB testing Frameworks and they're used for a lot of that kind of stuff and uh you know data
1:06:09
anal data scientists like to look and cut this data up and slice and dice to figure out like why metrics are going
1:06:14
down or like what is going on to get to like the root cause of things so uh that's important but keep in mind that
1:06:21
these metrics are not very often actually reported to Wall Street because like because of I gave a good example of
1:06:28
why okay so we have um uh derivative
1:06:34
metrics so um derivative metrics are great like I love them because they
1:06:40
really U illustrate like where the growth is headed for the company and
1:06:45
they're more sensitive to change uh I like the I like to use percent
1:06:52
increase instead of absolute increase because then you can get a better idea of like
1:06:58
the numbers that are going to happen and investors usually care about percent increase instead of absolute increase
1:07:05
and uh you know year-over-year growth that was the only number that Zuckerberg cared about that I worked on so uh that
1:07:12
was like the big one that he cared about and um it's funny so when I worked at Airbnb in 2021 we were working on
1:07:20
year-over-year stuff but we actually changed all of the metrics because of the pandemic
1:07:26
uh because like if we did in 2021 if we did year-over-year in March like March 2021 then it's like wow like we have 700
1:07:35

800% growth like we're doing great but it's like that's actually not what's happening it's just that like the year
1:07:41
before was terrible and so that's why these derivative metrics are more sensitive right because they are subject
1:07:47
to changes on both ends where it's like the like the previous year's Delta and
1:07:52
the current Year's Delta they both impact the the year-over-year number so you that is actually dependent on the
1:07:58
Delta of two numbers and as opposed to like the total aggregate is based off of just one number so that's why you double
1:08:05
the sensitivity when you do use year-over-year and so what we did was instead we did year over twoe so we
1:08:12
compared uh in 2021 we compared everything to uh 2019 and we kept we kind of did that for
1:08:19
a while for 2021 and then in 2022 we uh actually moved we migrated it back to
1:08:25
year over year and it was like we just did year over two year for a little bit and so uh because we just didn't want to
1:08:31
compare against when everything was shut down so uh derivative metrics they matter a lot we're going to be looking a
1:08:36
little bit at those in the um dashboarding session we have today uh dimensional mix uh dimensional
1:08:44
mix is great uh so a big thing here is like um for example like I know when I
1:08:51
worked at Facebook they cared a lot about like growing in in the de veloping world and they that's why they came out
1:08:57
with like I don't know if yall remember like Facebook Basics and like the internet.org where they're trying to give the internet to like people in
1:09:04
India and people in Africa and stuff like that because they want to get everyone on Facebook that was kind of
1:09:09
their idea and um because they look at the dimensional mix and they're like you look at like census and population data
1:09:16

and they're like wow we only have like 2% of the people in this country but the main reason for that and then then

1:09:21

Facebook's like okay why don't we have people in that country and then it's like oh not very many people in this country are actually on the internet yet

1:09:27

and so um obviously that's changed quite a bit that's changed quite dramatically since I worked there in 2016 and because

1:09:34

the the world in general from 2016 to now has become dramatically more online

1:09:39

across across the globe which is a great thing but uh obviously we need we there's still more Improvement we want

1:09:45

there we have to have all the memes I need to know all the memes I need memes from everywhere in the world it's

1:09:50

important to me um and uh this is going to be where you can also spot Trends and

1:09:55

populations uh that was one of the things that I noticed when I was working on Facebook was like oh like you could

1:10:01

spot a trend like oh Ethiopia shut off the internet and it's like oh we lost all our growth in Ethiopia but it's like

1:10:07

that Trend and it that fired a bunch of data quality checks and a bunch of data quality erors but it was like actually

1:10:12

real that's just like what the world was doing at the time and so um you can spot

1:10:18

trends like that and understand things right dimensional mix is really great for root cause analysis as well so say

1:10:24

you see a dip in a total Aggregate and you're like why is that number going down then you look at the dimensional

1:10:30

mix and you're like oh it's going down even more like in the US or even more in

1:10:35

iPhone users or like and you can kind of find like the the like the subset population that is impacted the most and

1:10:43

then that going to help you figure out like what is what is actually going wrong and like in the Ethiopia example

1:10:48

it was like oh we saw the total aggregate dip a little bit but then it was like 100% in Ethiopia and then
1:10:54
everywhere else we we seeing growth so uh that was a great example of a spot
1:10:59
where that can kind of happen so uh dimensional mix important remember dimensional mix and mix shift those are
1:11:07
words that you want to be aware of when you're doing your um visualizations and metrics and all that kind of
1:11:12
stuff um we talked a little bit about uh retention and survivorship uh in other
1:11:18
the class we're going to just talk about a little bit here um this is like the number of uh days um the the percent
1:11:24
left after a number of days like I don't know like what there's uh I think there's 41 of you in this um lecture
1:11:31
today and the number of people who actually um have survived this boot camp
1:11:37
right I think 41 is dramatically lower I think the number because we had 130 and
1:11:44
I think there should be 100 in here so I think like half of you or a little bit more than half of you died uh but um but
1:11:50
we still got 40 40 of you in here and which is impressive that y'all made it so far but um yeah the survivorship
1:11:57
analysis is important because it helps you see like the long-term value and like the lifetime customer value
1:12:03
lifetime like customer LTV is another very important um metric that your business needs to have congrats on
1:12:10
getting to the end of the day to lecture if you're taking this class for credit make sure to switch to the other tab so
1:12:16
that you get credit