Fact Data Modeling

Fact Data Modeling Day 2 Lecture
How Meta models Big Volume Event Data

The blurry line between fact and a dimension

Transcript:

1:32:24
is it a fact is it a dimension like how do we know what is the way to uh really
1:32:30
differentiate these things so one of the things that came up a lot at Facebook in
1:32:36
terms of just like these definitions that like they sound so similar but I
1:32:42
think they really illustrate the difference here so when I worked in growth at Facebook there is there was
1:32:50
two concepts you have dim is active and you have dim is activated these are two
1:32:57
different uh Dimensions that are on a user object so
1:33:03
for example dim is active was based on well did they have any activity I think
1:33:10
it was like did they show up for at least a minute on the app or they showed
1:33:16
up for less than that but they had an engagement like they had a like a comment or a share or something like
1:33:21
that they either like showed up for at least a minute or they uh engaged in the app in some way so but if the if the
1:33:30
dimension is based on that based on an event like they liked or commented or
1:33:36
shared something then is that really a dimension at that point or is it just an aggregation of facts and that's a great
1:33:44
thing to I'm trying to illustrate to y'all is it's
1:33:51
both right it's kind and and that's where where this uh whole modeling
1:33:56
exercise can get really really dicey so I just want to compare that with this
1:34:02

other flag dim is activated so I don't know if y'all have ever had needed to take like a mental health break from
1:34:08
social media and you decided like well I'm going to uh deactivate my Facebook
1:34:15
account and then what you do there is like you actually go into the app and you deactivate your account and then
1:34:20
that puts a flag on your account that you are now deactivating which is not an aggregation though right
1:34:26
because like it's just one thing and so a lot of times like that will change the
1:34:32
state of your account from activated to deactivated and then that is uh an
1:34:37
example of something that is a pure Dimension that one is going to be like because that's based that's an attribute
1:34:44
on your user object it's not based on the events that
1:34:50
you generate or anything like that so those two are very interesting because
1:34:56
technically you can um like at Facebook it was very confusing because there's
1:35:02
another way that Facebook can work where you can deactivate your Facebook account but you can keep messenger active it's
1:35:08
like you can you can go deactivate Facebook keep messenger active and uh so technically in that case you can have
1:35:14
dim is activated is false and dim is active equals true and those are like
1:35:20
the messenger only Facebook users cuz they like they don't like the feed but they still want to like have the DMS of
1:35:26
Facebook so uh just letting yall know that like these dimensions and fact uh
1:35:33
definitions like they do become blurry and that's an important thing to
1:35:39
remember so a big thing that can like kind of you want to think about uh when
1:35:44
you are like creating a dimension out of an aggregation of facts is well what is
1:35:52

the cardinality of that Dimension cuz a lot of times like you want to bucke tize it because it like for example the fact
1:35:59
that I put I had 17 likes on Facebook and then like say we wanted to look at
1:36:04
all the other people who liked the that made 17 likes on that day on Facebook
1:36:10
that's probably like the wrong bucket right it's probably maybe like 10 to 20
1:36:16
and then that's the bucket that we want to do like so you want to do this thing called bucketization when you are
1:36:22
creating uh your Dimensions out of your aggregated facts because then you can
1:36:28
kind of create smaller number of buckets and then that can make your group buys
1:36:34
more informative because uh a lot of times if you have extremely high
1:36:39
cardinality then the data is going to look really strange because like there might even be someone in that data set
1:36:45
who's like okay I liked 1,7 things today on Facebook because they were just like
1:36:51
just just going going and they were on Facebook for like 12 hours and they were just like scrolling like like like like
1:36:57
like like like like like and like but they're the only one in that bucket and so like that's where like maybe you make
1:37:02
a bucket that's like 500 Plus or 100 plus or something like that so you can
1:37:07
capture the long tail of people who are not uh like so that they aren't put in a
1:37:14
bucket by themsel because you very very rarely when you're doing these bucket
1:37:20
zations do you want to like ever have like only one one entity in each bucket
1:37:25
like because whenever you're doing Group by right there's kind of a an implicit
1:37:31
assumption when you use the word Group by and that's that you're forming groups
1:37:36
and if your dimensionality is so high
1:37:41
that like you have groups of one that's uh I don't know that that that doesn't

1:37:48
sit right with me and that's like can like it can interfere with your analyses and like a lot of times s that like
1:37:54
another thing that can happen is like if you have a group of one uh you come across this other problem of uh like
1:38:02
predic predictive power where like you don't you don't have normality anymore
1:38:07
if you have one data point or one user or one thing right you need to have at least like a couple like 30ish right
1:38:14
before you get that normality assumption in your data so the main thing I'm trying to say here
1:38:22
without getting too like data sciency and Technical about it is bucke tize
1:38:27
things into you know maybe like at the very most like 10 values like so that
1:38:34
you can see the different things usually five five is going to be probably The Sweet Spot 5 to 10 is going to be a
1:38:40
really beautiful range of like when you're bucketization
1:38:53
like is um needs to be informed right where like
1:38:59
you should be looking at the distribution of the data to see like oh what's where is like the this is where
1:39:07
like if you do like a a box and whiskers plot where you have like the median and then you have like the like you can
1:39:14
think of it as that way where you do like the you have like the the zeroth percentile to the the 10th percentile or
1:39:20
like the zero percentile to the 25th percentile 25th to 50th 50th to 75th and
1:39:27
then uh 75th plus and that would give you like four buckets and then maybe you slice it more where you do like 0 to 10
1:39:34
and then 90 plus and then 75 to 90 and like you can do it like with that case you would have six buckets where like
1:39:40
you'd have like Quin tiles or cor tiles or however you want to like slice up your data like but like it should be

1:39:46
informed from like actual like statistical distributions not just like
1:39:51
well like thought 17 to 30 was a good bucketization for this right and that
1:39:58
like you kind of like just um pulled it out of thin air so uh yeah don't do that
1:40:05
like because a lot of times like if you do that I've been there I've been there where I've pulled my bucketization out
1:40:10
of thin air and like then you like create all this data and then people are like so how did you define your buckets
1:40:17
and then like your answer is dumb and then people like are like wow like why
1:40:23
did you do it that way and then you have to like think about it and that's where like a lot of times though when you're
1:40:28
bucketization
1:40:37
that they want to do so like it's a trade-off though because if you add if you don't bucke tize then uh your
1:40:43
cardinality of that is or you're going to have an extra column and that the cardinality of that column is going to
1:40:49
be a lot higher and it's not going to compress as well so it's definitely a trade-off in terms of those things but
1:40:54
like generally speaking the flexibility of uh having the non bucketization
1:41:23
facts Dimensions they're blurry they can be based on each other so um because if
1:41:29
you think about it from the other angle like I I want to just talk about the dim is activated one more time there's also
1:41:36
uh the event of you going and changing that value which is an event that has a
1:41:42
Tim stamp where it's like you mutated your activated flag at this moment in time which is a fact but it's also a
1:41:50
dimension because it's like the the State the state is the dimension the
1:41:55

action is the fact so like you actually kind of have both in that case as well so just letting you all know the big
1:42:02
thing here is that like a lot of this the delineation between these things is kind of
1:42:07
blurry well uh like Dimensions um like
1:42:13
the big thing with Dimensions is they are the things that you group on like when you run like group by in your SQL
1:42:21
queries that the these are the values that are going to show up like user ID or country or device or gender or you
1:42:30
know I don't know like scoring class or whatever like and these can be both high
1:42:36
or low low cardinality right user ID would be considered like a high cardinality dimension whereas country is
1:42:41
kind of more of like a medium cardinality Dimension and then like gender is like a pretty low cardinality
1:42:47
Dimension um and then uh and the last bit for Dimensions like is they
1:42:55
generally come from a snapshot of state so at Facebook Netflix Airbnb the main
1:43:02
way that Dimensions come about is there's a production database and we we
1:43:08
take a snapshot of the production database at a moment in time and whatever those values are in that at
1:43:14
that moment in time are what the values are for that date um and so they come
1:43:19
from a snapshot of state that's like a very another important important thing to think about when you come come around
1:43:24
with Dimensions okay so uh facts on the other hand are going to be kind of the
1:43:30
opposite where like these are the things that you aggregate you sum or average or count or do all those things with them
1:43:37
like and those are going to be the big ways to like think about it where they go inside the aggregation function
1:43:44
whereas Dimensions go kind of outside with the group by so facts are almost

1:43:50
always higher Dimension because the number of things I do as a user of an app is usually more than one because I'm
1:43:58
one user who can do many things and I can log in I can delete my content I can watch a video I can you know you just
1:44:04
got to think about how many things that one user can do and it's going to be generally speaking a lot higher than uh
1:44:11
one thing and even if they can do one thing they can usually do it multiple times so the last bit that kind of
1:44:19
distinguishes facts and dimensions is facts come from law they are going to be generated like when
1:44:27
an event happens uh and a lot of times this is going to happen from logging so
1:44:32
that's going to be a big thing to kind of distinguish these things but keeping in mind that you can aggregate facts and
1:44:39
they can turn into Dimensions or facts can also change Dimensions I don't know if you all have heard of change data
1:44:45
capture uh change data CDC is a great example of the extremely blurry line
1:44:51
between the two where youve you kind of model A a state change of a dimension as
1:44:58
an event or as a fact and then you can kind of recreate your Dimensions at any moment in time based on the stack of
1:45:05
changes that have happened and so that the CDC is like way blurry because
1:45:10
that's like right I feel that that's like sits right in the middle between a fact and a dimension so yeah it's pretty
1:45:17
cool stuff so um when I worked uh at Airbnb I
1:45:22
wor worked in pricing and availability uh which is a fun place to be um and
1:45:29
there's there's some there's some things that I think are interesting about uh
1:45:35
price for example price as uh as an attribute on an Airbnb listing on like
1:45:42

so on a specific night you have a price for an Airbnb listing and is that price is that a fact or is it a dimension
1:45:50
right and it's one of those that seems like it might be a fact right because
1:45:55
you can sum it you so you can see like okay if I bought every night at this Airbnb for a month you can sum it up and
1:46:03
kind of see what it would be um so that's one thing to think about or you can average it you can count it it's all
1:46:10
like that prices are doubles it's kind of rare for like a double or like a a decimal number to be uh Dimension um one
1:46:19
of the things that's interesting about this is that it kind of is a Dimension though because of the fact that it's
1:46:27
it's the it's the attribute of the night where it's like this is just the price of that night it's a state right so
1:46:34
let's go to the next slide there's two things to think about here uh you have
1:46:41
the a host can go and change their settings where they might offer a discount like a last minute discount or
1:46:48
like a early bird discount or something like that and when they change their settings that will log an event and that
1:46:55
is a fact because it's logged uh but price is actually derived from all of
1:47:02
the settings that the host has set and these and these settings are State and
1:47:09
uh since their state price is actually a dimension and so it feels like a fact
1:47:17
but it's actually a dimension and that's always one of the things that I like was like wow this is this is like um kind of
1:47:24
a weird thing uh about like how these kind of different elements of the puzzle
1:47:30
of data engineering kind of fit together and I always thought that that was so strange okay let's talk a little bit
1:47:36
more about like uh these Dimensions that like are based on
1:47:43

facts um we talked about dim is active right that makes sense you also have like dim bought something maybe they
1:47:50
they ever bought something right uh if they're someone on the website and they
1:47:55
that's a dimension that's based on one event that means that they they have not zero events in their fact uh purchases
1:48:03
table or whatever it's a rollup uh you also have things like dim has ever booked so um these are great where it's
1:48:11
like okay did they ever show up did they were they ever labeled fake that was when dim ever labeled fake that was a
1:48:17
dimension that I had at Facebook when I worked on the fake accounts table and that one was essentially like a feature
1:48:23
where it was like a lot because a lot of times like if an account is ever labeled fake that is a very strong indicator of
1:48:31
like how that account is going to behave over the long run even if it gets unlabeled fake so that is a it can be a
1:48:38
very good Dimension to bring in but again it comes back to okay how do you get dim ever labeled fake that is comes
1:48:45
from the fact that like at some point uh a machine learning model created a label
1:48:51
that said this account was fake and um and so these these ones usually only ever flip from false to True um you can
1:48:59
think of these uh you can think of Dimensions that are a little bit different like uh you could say like dim
1:49:04
is monthly active where it could flip back where it's like okay you're active one day of a month and then you're
1:49:11
inactive for 31 days and then so it flips from true and then back to false so if it has like kind of another time
1:49:18
component to it you can have that as well that's another very powerful thing to think about um so then you have days
1:49:26
sense uh this is something we're going to be covering in uh week five of the

1:49:31
analytics track we're going to go a lot deeper into this um kind of it's like this retention analytics uh cohort curve
1:49:39
stuff that will be very powerful they use this a lot at um Facebook you can
1:49:45
think of it as like okay everyone who signed up for the app on this day how many of them are still active in a week
1:49:51
in a month in a year and you can kind of see like usually it starts off very high because on the first day everyone showed up so it
1:49:58
starts off at 100% And then it kind of slows down and then it like it kind of reaches what's called an ASM toote where
1:50:04
it's like there's like 10 or 20% of the people left who are going to keep coming back and those are like the sticky
1:50:09
people and uh if y'all want to like look up more about that to be ready for week
1:50:15
five look up a thing called J curve or retention analytical pattern and these
1:50:20
are also based on um um these are even one layer on top of the aggregation
1:50:27
right because it's like okay you were active on one day and then it's like how many days has it been since you were uh
1:50:33
since you were active on that day or how many days has it been since you signed up and so then you can see like okay
1:50:38
what is what does that curve look like and so and that's like even one more layer on top it's like a dimension on
1:50:45
top of a dimension which is based on the the aggregation of a fact so it gets a
1:50:52
lot of this stuff gets very muddled that's one of the big things I'm trying to illustrate to y'all is that but um
1:50:59
that's a good thing to remember here is a lot of we have a lot of these dimensions and in our lab today we're going to be looking at dim is active
1:51:05
it's going to be one of the big ones that we look at so okay so let's talk more about uh
1:51:13

categorical uh fact bucketing of Dimensions right so we had the simple one we had a very simple bucketing
1:51:20
example in week one where we bucke tized users based on uh points right but a lot
1:51:28
of times it might be more complicated than just like one value where it's like points like for example for Airbnb super
1:51:34
hosts there's like there's like a a bunch of criteria you have to look at a bunch of columns together to create that
1:51:41
um Dimension uh the you have to have like a certain amount of ratings they have to be a certain number you have to
1:51:47
have a certain number of bookings right and you have to have a certain amount of Revenue or something like that there's like a a certain number of uh to look at
1:51:54
maybe three or four columns to determine if you're a superhost at Airbnb and so uh like it's not always like you just
1:52:02
bucke tize based on the value of one column so a lot of these things can be more complicated than that like to
1:52:09
create these kind of more conditional columns that you can look at that are also very powerful and they can be used
1:52:16
for your analytics and you can kind of see like oh this group is very powerful versus like these other groups and then
1:52:22
you try to get more people into that group or like a lot of times these Dimensions can end up setting like the North Star for the company where it's
1:52:28
like okay we want to get as many people to fit this definition and you know at Airbnb that's
1:52:36
one of their things right is they want to get as many super hosts on the platform as they can but they don't want
1:52:41
to like like make that they don't want to change the definition of that Dimension because like that would defeat
1:52:47
the purpose right because they could just game the system and be like everyone's a super host right and then that works but they have to like come up
1:52:54

with some sort of criteria to fit people in there so that it's a meaningful Dimension and not just like gaming the
1:53:00
system so that's another thing to think about is that a lot of times when you make these uh Dimensions one of the
1:53:05
things that can be tricky is um they can be very hard to change uh like um for
1:53:12
example uh there is another uh like one one of the other things that I was looking at when I worked at Facebook at
1:53:19
one time is like I don't know if y'all know but there is a hard limit on the number of friends you can have on Facebook you can only have 5,000 friends
1:53:25
and um if you it won't let you add more than that and uh because they realized that like well if you increase that
1:53:32
limit like what like you can't it's a one-way door right it's like once you
1:53:37
increase it you can't go back you can't like force people to be like oh like you have 5,000 friends and or you have 8,000
1:53:44
friends and we're going back to 5K so you have to like unfriend 3,000 people so like what is some of the times like
1:53:50
when you have these mentions they can have a very strong anchoring effect on
1:53:56
the product like LinkedIn has the same same exact problem and it's one of the things that I hate the most about
1:54:02
LinkedIn so LinkedIn has a a connection request limit as well it's a 30,000 and
1:54:08
uh once you hit 30,000 you can't connect with more people which I always think is so so silly where it's like why is it
1:54:15
30,000 like why can't it be like whatever number come on Microsoft you're a trillion dollar company I'm pretty
1:54:21
sure you all configure out how to deal with people who have more than 30,000 connections but um anyways that's kind
1:54:28
of the idea is bucketization and dimension definitions like a lot of
1:54:35
times once these are set it's hard to change them I know another example for me when I was uh

1:54:42
working uh at Airbnb and this is one of the biggest impacts I had at Airbnb was
1:54:48
when I was working there uh we had a column called dim is available which is
1:54:55
can uh this Airbnb like is this Airbnb available on this night and one of the
1:55:02
things that's weird about that is that the old definition of that was kind of
1:55:08
did a host set a rule that blocks this KN that's the that was the old
1:55:14
definition and uh one of the things I did was I made a slight change to that which was can a trip be booked that
1:55:22
contains this Knight so those are slightly different definitions and like that was a I mean that just making that
1:55:29
change was a twoyear undertaking at Airbnb just to make that change even though the those definitions are almost
1:55:35
the same they're about three or 4% different because of like some edge cases like one of them being sandwich
1:55:41
nights which is where if like a host says okay you have to book at book at least three nights then um if there's a
1:55:47
reservation in two days then technically tonight is unavailable because you have to book three nights in a row and so
1:55:53
these nights are like squished in and they're unavailable whereas in the old definition they were marked as available
1:55:59
so like it's a great example of like when you determine the definition of a
1:56:05
dimension changing that especially if it's used throughout the company can be
1:56:11
very painful and can take a long time and so that was one of the big things I worked on was making that change and
1:56:17
getting the new definition of availability adopted throughout the company and that was is uh a fun problem
1:56:23
but also a massive pain like I was like holy crap like why is something so simple so annoying to deal with so

1:56:31
definitely uh when you are thinking about these definitions get as many people involved don't just like go with
1:56:39
something and be like done like really try to think about how these impact the business and that can make a big
1:56:44
difference on uh how many times you have to change them because know that changing these Dimensions is not cheap
1:56:50
it's actually very expensive so so uh big thing to remember when you were doing your categorical Dimension
1:56:58
values well um here's uh another great thing to think about is um should should
1:57:06
you use Dimensions or facts to analyze users like so do we care more about how many activated users there are or how
1:57:12
many active users there are and I mean like the answer is probably both uh
1:57:22
think that like there's an intuition here probably from most of the people in this presentation that active users is
1:57:28
probably the more valuable of the two but activated also matters because you
1:57:34
can especially like if you and a lot of times you might use both of these together because imagine if you have
1:57:39
active users divided by activated users so you can see the percent of users who
1:57:45
signed up who are active that's a beautiful analysis right that's another
1:57:51
great metric that you could look at that I think could be a very powerful metric right so the thing that you want to
1:57:57
think about here is what's the difference between like signups and growth right signups versus growth those
1:58:03
are going to be the two kind of things to think about when you're kind of going through this process here so those are
1:58:10
going to be the two uh like so it really depends on what question you're trying to answer so they both are very valuable
1:58:16

when you're doing um analytics Okay so what we're going to
1:58:22
talk about now is this is going to get kind of Technical and um I I I hope
1:58:29
y'all uh can can get through this so um at Facebook they were trying to
1:58:37
figure out a way to store historical growth information like because one of the very common questions that are is
1:58:44
asked at Facebook is how many monthly active users do we have weekly active users daily active users do we have and
1:58:51
uh one of the things that uh is annoying about monthly weekly daily active users
1:58:59
is um especially monthly monthly is annoying because if you think about like
1:59:04
if every day you needed to process like okay is this user monthly active then you have to look at the last 30 days of
1:59:11
fact right and then do group Buy on the last 30 days of facts and then every day
1:59:16
you have to look at the last 30 days of data and like even though 29 of the 30 days haven't Chang changed only one of
1:59:23
the 30 days has changed so like if you think about it in a very like uh naive approach that would be the very naive
1:59:30
approach right is like just every day process 30 days of data and do a group Buy and see like okay they were active
1:59:36
on at least one of the last 30 days um I don't know uh I find that
1:59:43
unsatisfactory I think that most of y'all like as a data engineer especially at Facebook if you think about that like
1:59:49
yeah there's two billion users and then uh you think about the fact data there where two billion users each user does
1:59:55
50 rows a day so that means you have um uh that's going to be uh a 100 billion
2:00:03
rows a day of fact data and then if you want to do monthly active you do 100
2:00:08
billion times 30 so then you're uh in what three trillion three trillion rows
2:00:15

so if you did it that way right where you're looking at three trillion rows every day um oh don't know that sounds
2:00:22
expensive right and that sounds like not the most efficient way to do things so
2:00:27
they used uh the cumulative table design just like um we talked about in week one
2:00:33
we did cumulative table design there and what we do is for cumulative table
2:00:38
design we are creating new um a new
2:00:43
table that we we process 30 days and then the then when the next day comes in
2:00:50
we drop off uh the 31st day and then we keep everything else and you could think
2:00:55
of it as like kind of a naive approach to this would be you would have a user ID you have the current date and then
2:01:02
you have dates active where dates active is an array of all of the days where
2:01:08
that user was active and you can cumul cumulatively um add to this array so
2:01:13
every time you see a value you can add like another value to the array so that you can see like oh this user was active
2:01:19
on this day this user was active on this day da da da um one of the things that sucks about
2:01:25
that though is then you have this big old array of dates and like you don't really need them because dates are
2:01:31
mostly uh you can think of them as mostly As like an offset so think about
2:01:37
it this way instead like if you had it where you have this thing called a DAT
2:01:42
list int so instead like imagine it this way where we have user 32 January 1st
2:01:48
and then this structure here is going to be um the the number of days uh like when that
2:01:56
user was active so see this one here that means they were active on January 1st right and then you go back right
2:02:02
what this is going to be um 1 two 3 four five six 7 so here they're going to be
2:02:08
seven days back right so this is like I December uh 23rd or 24th or 25th or

2:02:13
something like that of the of the previous year they were active and then you can keep going back so every every
2:02:20
day in here is or every bit in here is one day back and then you can get kind
2:02:26
of a history of all the users and what days they were active and so that can be a very powerful way to see okay this
2:02:33
user was active on these days and this could be a very powerful way to store this data because now you can store 30
2:02:41
days of history as an integer data type and if it's an integer data type
2:02:48
you uh you get very good compression in ERS are some of the highest compressing
2:02:54
data types that you can work with so that's going to be a big thing that we're going to be working on today in
2:02:59
the lab is we're going to be going over how to create this beautiful awesome
2:03:06
data structure called the DAT list int okay so that is that is actually the
2:03:11
last slide for the presentation today congrats on getting to the end of the day two lecture dat lists are crazy
2:03:18
right I'm excited for you to check out the lab as well if you're watching watching this on the platform make sure to skip over to the next link uh on the
2:03:26
platform so you get the credit that you deserve here thank you so much and keep learning one of the tables that we have