Data Quality Patterns

Day 1 Lab

Transcript:

45:14
you can get the time tracking that you need uh for the lab I'm excited for you to check this out so in this case uh I

45:22
already started building out something here like a lot of times you want to start out with like some sort of nice title

45:27
I'm calling this the exactly Inc user growth Pipeline and then what uh you

45:33
know we talk about like we want to measure website traffic and user growth and it's kind of the idea is why it's

45:39
called the user growth Pipeline and then what we're going to be doing is talking

45:45
about okay and they say the goal of this pipeline is to answer the following questions so this right here like you'll

45:53
see as if a data analyst comes to this spec and then they're like oh I need to know about this or oh I need to know

45:59
about this metric or I need to know about this it's great you can definitely uh like get a lot of information just

46:06
from the description of a pipeline and obviously we have a couple more things we want to kind of go over here um

46:12
hopefully we have time here we I think we should so when we're going over like business metrics one of the things that

46:19
I think is important is this is going to be like a table so let's go ahead and insert a table here um and I think

46:26
there's going to probably be uh it's going be probably a 3x4 table here so in this case we have like metric

46:33
name and then in this case we have uh I like to call this like uh is guard rail

46:40
I'll explain what that means here in a second and then uh what we will uh then

46:46
it's like this one is like is Ratio or something like that or like no no my bad
46:53
this is it's it's like this so you take is guardrail move him over here and then
46:59
this is like definition um so in this case we want to come up with some Metric names uh that
47:07
that are coming from some of these questions right so one of the things you can think about is okay what percent of
47:14
traffic is converting to signing up so I think that's a good one so we could say
47:20
like sign up conversion rate that's probably uh a good metric I
47:27
like to bold the metric names um so the definition here right is going to be
47:32
like count signups divided by count website hits
47:39
something like that and um obviously this definition is like kind of sloppy
47:45
it would be like as you kind of Define more of the schemas uh further down line that is going to be one of your things
47:51
so what do I mean by is guardrail so a guardrail metric is a metric that
47:57
signifies um uh a problem to the business that is
48:04
really bad so like if you have a a metric that is a guardrail metric what
48:09
that means is it's protecting uh the business of or or like the measurement there is trying to
48:15
protect the business like especially in the case of like an AB test so imagine if you launched an AB test and the ab
48:23
test essentially says um oh this AB test is very negative
48:29
on this metric and the metric is I don't know maybe it's something like profitability or it's like growth or
48:36
some other kind of really important metric and like essentially what the guard R metric will say is like don't
48:42
launch that test don't don't roll that out to production that means that that
48:48

uh and that test shouldn't run at all right whereas a non-g guard rail metric might go down but that's like more uh

48:56

it's not like whether that's good or bad for the business uh can be um it's less

49:03

clear and so that's where you can have other metrics that are going to be like not guardrail metrics I think sign up

49:10

conversion in this case is yes is a guardrail metric right so what about uh

49:15

other ones here that like we can think about uh okay what I like this one too what percentage of signups uh convert to

49:22

paying customers so I like that one this is going to be like um purchase conversion

49:30

rate and in this case uh oh yeah let's we can make this like a little bit like that so in this case it's going to be

49:37

count purchases divided by count

49:42

signups that's going to give us that this one I think is also a guard rail for a different reason so I think this

49:49

this first one is a guard rail because it indicates that there's a problem with the landing page if this number goes

49:56

down um and then this one is uh a guard rail

50:01

because it indicates that there's a problem with the checkout page if this number goes down now let's uh let's

50:07

create one more metric here that maybe isn't a guardrail metric that's uh maybe like one of these like where are these

50:13

people coming from so I think in this case like we can say um it's going to be

50:20

like uh traffic breakdown or we say yeah tell you traffic breakdown call it like

50:26

that maybe and then in this case we're going to have like count website hits

50:32

Group by uh referer I'll explain what referrer here is in just a second that's

50:38

essentially like think like referrer as like if I am on LinkedIn and then I click on a link on LinkedIn and it goes

50:44
to exactly.com then LinkedIn is the referrer it's like where you came from
50:50
so that's kind of the idea here would be like okay we want a breakdown of like all of the
50:57
the websites but grouped by referrer right so that' be the kind of like it's almost
51:02
like that's kind of the idea here uh obviously like this metric like you could have this metric be broken down
51:09
into like traffic LinkedIn percent traffic Twitter percent and you can kind
51:15
of like pivot it out and that's a I mean and honestly that's probably like if if I wasn't trying to build this spec in 40
51:21
minutes I would do it that way because then that's just one number whereas this could be this is like dozens of numbers
51:27
right because the number of websites that point to my website is pretty high so in this case I want to say this is no
51:34
because a couple of reasons like one is an AB test is not going to impact whether or not my traffic comes from
51:41
LinkedIn or Twitter or wherever like that's all on me and how I change my
51:46
social media channels so this is not a guardrail metric because of the fact that like whether like it like whether I
51:52
do an AV test on my website should have no bearing on where the traffic
51:57
so this is kind of the idea behind like when you're coming up with business metrics and how they could be useful in
52:05
your kind of diagrams right so we got uh a couple more pieces here obviously
52:12
there's a lot more different business metrics that you could come up with uh we're going to we're going to stop and put it there uh but we could also come
52:19
up with some more later on but let's uh let's put an enter here I I don't like have like the flow diagram be on like
52:25
two pages like that that but um so in this case let's go to Lucid chart and
52:31

kind of go over how we can build a flow diagram of this
52:41
so okay can I just like copy this guy
52:46
because this guy was already so good okay great we can just start with
52:52
the one that I uh used for the presentation today um so let's go back
52:57
to uh kind of the business requirements here and you'll see uh it says let's
53:04
let's like kind of look for some things here that I think are important to call out so you see this like what is the
53:10
geographical and device breakdown of that traffic so what that means is we
53:17
need to figure out somehow figure out where the traffic is coming so uh traffic generally speaking is
53:27
uh how we measure that like where it's coming is there's an extra step in that
53:32
process there and that's going to be what's called like IP enrichment where you need to like take the IP address and
53:39
then pump it to some website and then it will tell you like oh this is a California IP or this is a New York IP
53:46
or this is an India IP so we need that but we also have this device breakdown
53:52
which has its own thing right where a lot of times you can have a device identifier or you can have user agents
53:59
so there's going to be another layer here probably there's probably two layers here there's a uh an IP
54:05
enrichment layer and a device enrichment layer that we're going to need to include in our source data because it's
54:12
not in the source data like with that being said let me kind of just go over what's going on with that so uh so if we
54:20
go here we say like let me close that so we say like select star from events this
54:26
table is like the raw the rawest of the raw data you see you have like an event
54:32
ID and then you just have like this big old blob a Json it's kind of terrible um
54:38

but this is the um this is essentially the data that we're going to be working with and this is the RW data so this is

54:47

our source schema but in the specs generally speaking you don't have to put

54:52

the source schema in the spec right the source schemas do not matter it's all about uh the the data modeling that

54:58

you're doing specifically because people don't really care about like where the data came from they kind of trust that

55:04

you're going to do the right thing but uh this is I'm just trying to give you an idea of like what's going on here in

55:11

here like we have that table right so in this case it's just a logs table and

55:16

we're just going to call this um we call it log or we're going to call this events and uh one of the things that's

55:23

different about this uh schema and um another schema right is

55:30

that we actually aren't going to be doing any joins with Dimension tables because my business is pretty new so I

55:36

don't have any like Master data but we have things in the middle here that we want to also include so a lot of times

55:45

those things can be modeled with like a square I like squares here so um what we

55:51

want to do here is pop this guy here and then uh move this guy here and move this

55:57

guy here and then like I kind of want to move all this like over and then move

56:03

him oh just move all of it over move this guy up here this guy up here this

56:09

guy up here there we go so now that's like a that's like a a little bit better looking thing I mean some of this stuff

56:14

is probably now too far over but so this first one here is called like IP

56:20

enrichment and then this second one here is called user agent enrichment and so

56:26

we you can kind of go over like what those things are doing but those are mostly API calls to take the data that

56:32
you have and then make it like more complete what's going on sometimes a
56:37
lucid chart there we go that's more like what we're looking for so then we have our events that create our kind of this
56:46
will be our highest level table that you can think of this is going to be like if you go back into here uh you see how we
56:54
talk about what is the geographic device breakdown of that traffic so now we have
57:01
that that's going to be our first table but this is going to give us both though right because every event in that table
57:08
there's events for website traffic there's events for sign up there's an event for purchase all of these are also
57:15
in there those are all different types of events that are in this events table
57:20
so then what we have is we have a table here called um I would call this one maybe core and we can call this fact uh
57:28
website events or like maybe yeah call it fact website events I'm I'm I'm happy
57:34
with that name uh let's make this a little bit smaller so it's like actually fits in the Square um so this is one
57:42
table that I think is really useful uh there might be some interesting columns in this table that we want to consider
57:50
and uh that's what we're going to kind of do next here like we're going to and this is going to be another beautiful
57:56
part of like kind of showing you how these kind of specs are generally
58:01
generated is they have uh like you kind of go back and forth between like the
58:07
schema page and the flow page and then kind of trying to understand what you're what you're going after so we're g to we
58:15
have the flow diagram but let's let's think about the schema here so in this case we have um we have that table right
58:22
we have core. website events let's make this a little bit

58:28
smaller and then a lot of times like you want to like do a description here oh let's make this even smaller let's make
58:34
this like a like 14 point there we go and then we want to talk about how
58:40
uh why is this not bold this should be so usually the table names are bolded
58:46
and then you could say like this table is a list of all events for exactly.com
58:56
and includes uh IP enrichment and user
59:01
agent enrichment for country and device
59:08
specific information so you can kind of think of like you can add a little like
59:13
description a little bit of flavor if you're going to uh talk about the table that you're creating and so then what
59:19
you can do right is you can go ahead and webly want to add like a table here
59:25
usually this table is going to be uh you need three columns I'm going to put probably we'll
59:31
do 3 by8 so in this case you why is this so
59:38
big we go 14 so we have column name then
59:43
you have column type again it's freaking massive column type and then you have a
59:50
column column comment there we go column comment so uh
59:58
and then it's like why does it keep doing that can I like just like can me
1:00:04
like there okay so um in this case we want to
1:00:09
think about all of the possible comment or columns that we want to add in this
1:00:15
table and this is going to be like kind of a quick and dirty kind of fact data modeling exercise so um one of the
1:00:24
things that I think is important here is uh the website events table May and
1:00:30
may not have a user ID in it so let's go ahead and add user ID first and user ID
1:00:37

is type big int and this is um we want to talk about okay so why can it may or
1:00:43
may not have a user ID so if you're logged into the website and you like go
1:00:48
look at a course and you click a button that your user ID will be tied to that
1:00:53
event but if you're logged out you're like just checking it out like you won't have a user ID you're you're you're like
1:01:00
logged out so um generally speaking so we can say this column is nullable for
1:01:06
logged out events um this column indicates the user who made this who
1:01:18
generated this event so um yeah this this still too big this is should be
1:01:24
like 11 okay um then uh we can think about other ones we have country right
1:01:30
and country is going to be a string and the country
1:01:35
um uh the IP the country associated with
1:01:41
the IP address of this request then you have like a device
1:01:48
brand like that string the device the brand associated
1:01:56
this request right some of these column comments are pretty whack because like
1:02:02
they're because they are obvious like because the column name is pretty self-explanatory so this is where um you
1:02:09
want to be careful with some of these names I would actually throw in dim here probably do dim device brand and dim
1:02:16
country and then um then you have another column here uh I think you have
1:02:21
uh action type uh this is going to be a string uh
1:02:27
is an enumerated list of actions that a user could take on this website uh and then
1:02:36
you might want to add here like sign up watch video
1:02:41
uh like go to landing page you know Etc
1:02:46

like it's like an enumerated list and like that's where we can explain to people that there is only a certain

1:02:52

number of values here and so that can be one then we like I think we have only like three more columns to kind of like

1:02:59

bash out here I think you have event timestamp and this is a Tim stamp and

1:03:05

this is uh the UTC timestamp for when this event occurred make sure to always

1:03:12

log your time stamps in UTC but this is a great example of why documentation matters even for time stamps because you

1:03:20

when you have a time stamp like is it London time is it East Coast time Pacific Time who knows right so this is

1:03:26

a great way to kind of cover your bases when you do have a time stamp because you're like yo this is obviously a UTC

1:03:34

timestamp because I'm labeling it as such so then I think you have another column here uh you have like other

1:03:40

properties is a column and I call this a map this is going to be a map of string

1:03:46

string and this is going to be uh any other valid

1:03:52

properties that are part of this request so this column is going to not be used

1:03:58

as much and then we have one more column here DS this is string and this is this

1:04:03

is the partition column for this table

1:04:09

so that's kind of the kind of uh structure so far I totally forgot host name let's let's go ahead and add that

1:04:15

like so we're going to put like dim um host name right this is going to be string up what is the host associated

1:04:23

with this event exactly Zack wilson. Tech Etc right thank you

1:04:32

that is totally right yeah that was that was one that was missing in this right see and this is why you build a spec

1:04:38

right because now like we're working on this together and like now we already caught something and that's when you have two sets of eyes so there's

1:04:43

actually a we have a logged out user ID and this is going to be a big int and uh

1:04:50

this column is a hash of IP address and device inform

1:04:57

so in that case like essentially what we say is and this is what they use for the primary key of this table right since

1:05:02

user ID is nullable you're totally right uh but like we can have both here so that we can see a link between the two

1:05:08

so in this case the logged out user ID is never knowable because every uh Network request has an IP address

1:05:14

associated with it so in this case what we say is for every logged out user ID and so it's like if if if a request

1:05:21

comes from the same IP address at the same time then we could we can consider that a duplicate if they're like at the

1:05:29

exact same time but obviously event Tim stamp is like down to like Nan seconds potentially right so like the odds of

1:05:35

that happening are pretty low but um so that's a good thing that you can add right so especially when you're in this

1:05:40

kind of logged out world where like you don't have anything to no ID that you can enforce like an entity on a lot of

1:05:47

times you have to create your own ID out of the data that you have available and so a lot of times they use I don't know

1:05:54

y'all have ever heard of a um murmur 3 this is murmur 3 hash right we'll use

1:06:01

the murmur 3 hash right that's a cool hash by the way I just like I like murmur 3 it's one of my favorite hashes

1:06:07

um so uh that's totally right thanks for that feedback on like how to do this table so um then we can we can probably

1:06:14

add that in here the um the unique identifier for this table is logged out

1:06:21

user ID and event time stamp so uh that can be useful um for this
1:06:30
kind of to understand stuff right so we have our schema we have our logged out user ID we have all this stuff and so
1:06:38
this is going to be one table in like all all this thing right and so but this
1:06:43
table is like pretty much like we only need probably an aggregate table as well and I think that that will give us uh
1:06:51
like pretty much all the schemas that we need and so uh let's go back to this uh
1:06:56
Lucid chart here and we have this a here uh we're going to call this um core a
1:07:02
website events we're going to essentially just call it the same name as kind of the other
1:07:09
one so then uh these uh let me move these over a little bit okay why is this
1:07:15
one like so funky okay whatever there so
1:07:21
um a website events is going to be one where we kind of uh add some more uh
1:07:26
layering to it right where we have uh in our kind of let's go ahead and copy like
1:07:33
like a little bit of this table and go down here and then um in this case we're going to say um core. a website events
1:07:41
this is like way too big make this like 15 and fold it and then we can say uh
1:07:49
this table is an aggregated view of all um
1:07:56
website events right so in this case we actually probably won't have a user ID at this point because it's going to be
1:08:02
aggregated up um it's probably going to be like maybe a daily aggregation or something like that uh so um because
1:08:10
that's usually what metrics you want like these comments and likes and all this stuff are going to be like daily Aggregates so uh I actually think that
1:08:18
this is close let's just get some of these uh so we have like sign up conversion rate right and then we
1:08:26
have purchase conversion rate there
1:08:32

like traffic breakdown so this is probably pretty
1:08:39
close uh that this is probably where I'm I'm feeling pretty satisfied with like taking a screenshot of this bad boy and
1:08:45
throwing him in throwing him into the actual dock here so we can uh kind of
1:08:51
see how this will work um let's go back down and finish this schema real quick
1:08:57
and then we're going to shift gears to the other pieces of this that are not included quite yet so I think in this
1:09:03
case the First Column here is going to be your uh kind of string right which is going to be your date like but a lot of
1:09:10
times there's actually we're not going to put that first let usually the last column like the partition column is almost always the last column so we'll
1:09:17
start with uh in this case there's probably going to be action type which is going to be a string and this will be
1:09:24
uh the enumerated action type um and then we talked about the
1:09:31
enumerated up here so obviously we could copy this column comment down if we wanted to or not and then in this case
1:09:38
we have probably country string this dim
1:09:45
country this this dim action type as well move this one
1:09:51
dim okay so uh countes
1:09:56
for the IP for the IP country um so we're going to insert like
1:10:05
a couple more rows here so in this case what we're looking for is there's going
1:10:11
to be an interesting column here that I uh I think yall are going to think is kind of an interesting one that at least
1:10:17
I like for aggregate tables but so we have country and we have like dim device brand right which is a
1:10:24
string device brand some of the column comments can be terrible like that but it's like what else do I put here like
1:10:30

you might put like Android iPhone Etc or however you want to
1:10:36
enumerate it out right and then uh in this case we have uh maybe uh so this is
1:10:41
where like it gets a little bit tricky or maybe we have event hour I'm going to put an integer here and I'll explain why
1:10:47
we're putting hour in here and I think it will make more sense so the hour this event took place um
1:10:55
in UTC and then in here we're going to say um M total events this is going to be a
1:11:02
big int this is the total number of events for this um slice and then we're
1:11:10
going to have two more columns here um insert row below then in here we're going to have aggregation level and I uh
1:11:17
I'll explain how this works and this is going to be a this will be a string and then we have DS which is a string which
1:11:23
is this is the date partition this table and then aggregation level is this is
1:11:30
how this a table is grouped um and then uh we can say values include uh so in
1:11:40
this case we have uh dim action type then like there might be dim
1:11:46
country uh dim action type so like you can put like a bunch of
1:11:52
different uh aggregations in here like and maybe have over overall um overall
1:11:57
is an aggregation level and so like and then there's probably one more with event hour right and so you can have
1:12:05
like all of these different aggregation levels in one table in one aggregate table so then you can like all the you
1:12:11
don't have to run the group buys you just have to say like select from this table where aggregation level equals
1:12:18
this so this is probably a little bit confusing to y'all but um if y'all like
1:12:24
we're going to be working a little bit more on this stuff in the applying analytical patterns uh week uh for the
1:12:31

analytics track we're going to be going over a lot more of this in detail of how these aggregation levels work but just
1:12:37
think of it as like what is the group buy here right and then you could say like dim action type put all of them
1:12:44
here dim dim action type uh dim country dim
1:12:50
device brand event hour there might be like
1:12:55
that one right this is like essentially like all of them together and then you have like all the
1:13:02
different ways that this could be grouped right could think of it that way as like um because we want to do this so
1:13:09
that when we are getting to the metric generation layer all we got to do is Select stuff for the most part we select
1:13:16
and divide and then that is how we can like so then the metric layer doesn't have to do any heavy lifting at all like
1:13:23
we do all the heavy lifting in Sp spark and we do it all in like the big data layer and then we don't do anything here
1:13:31
so that's kind of the idea behind uh schemas uh but what about quality checks
1:13:37
so we need to do quality checks on both of these tables and uh let's let's go ahead and look at this first table here
1:13:45
and like usually you put the quality checks at the bottom here so we quality checks again like way too big like 15 um
1:13:53
and then in this case uh we probably want um a couple here uh
1:13:59
like not null checks on um dim host name
1:14:05
dim action type event Tim
1:14:11
stamp uh dim country yeah logged out user
1:14:19
ID uh but um device brand could be null uh because
1:14:28
of the fact that and we can probably put that in here um could be null because of
1:14:34

bots that don't have a brand because like Google bot could hit my website and
1:14:40
uh it would log a record but like uh the Google bot isn't iPhone right it's it's like a weird like crawler thing so you
1:14:48
can think about that and then we have like um make sure no duplicates on
1:14:53
primary key obvious one uh then um there's probably
1:15:00
a dim host name is well formatted dubdub dub.
1:15:07
xxx.com right something like that all
1:15:14
right we'll put x.com to O OD to Elon um
1:15:19
so uh what else we got here you can see how there's like a bunch more here we could also say like row uh row count
1:15:26
checks um group on uh dim host name and check
1:15:34
week over week impr uh week over week uh counts for www. exactly.com
1:15:42
and dubdub dub. Zack wilson. te because those are like kind of critical Dimensions those dimensions are way
1:15:48
bigger than the rest of them and so you can have r r count checks there uh let's see here um there's en
1:15:57
enumerate yep there's one there's a couple more you have enumeration check on dim action
1:16:04
type should be sign up
1:16:11
purchase login Etc um so those ones are pretty obvious
1:16:19
there's there's going to be other ones as well that you could think of probably uh so I like that for the most most part
1:16:26
um I think let's do some quality checks for the second table real quick um so
1:16:33
like 15 quality checks uh in this case uh we have not
1:16:38
[Music] null and this is going to be on uh you have dim country dim action type but see
1:16:47

one of the things that's interesting here is like if we did the not null check on the table above do we need a do

1:16:55

it again right that's one of the things that's always interesting to me when I think about quality checks is like if

1:17:00

you're already reading in trusted data do we need to do these again so in some ways you kind of don't right because

1:17:05

these are not going to catch anything because if they're going to be caught they're going to be caught Upstream right um uh so like let's actually not

1:17:12

put those in there because like we don't need those right so we at row cap checks right um

1:17:19

overall rollup should have more data than any other rollup

1:17:25

um and then uh probably other things on uh things like what are some other ones

1:17:33

here that are going to be interesting event hour probably has some things right uh event

1:17:40

hour should uh look more seasonal or should

1:17:46

look like it's old seasonal pattern maybe something like that like it's like a advanced kind of check um

1:17:55

yeah M total events events uh should be greater than

1:18:02

some minimum number assuming that like we assume that there's going to be data

1:18:07

every hour or something like that right there's like obviously there might be a

1:18:12

time when there is no data so then we have our quality checks so then we have our we have our flow we have our flow

1:18:19

chart we have our table we have our quality checks and I think that's pretty

1:18:25

much it that like from like what we need to like get rolling with this pipeline

1:18:31

congrats on getting to the end of the SPEC Building lab I hope you're more knowledgeable about how to build good documentation because your stakeholders

1:18:38

will greatly appreciate it over the long run make sure to like comment and subscribe and share this with your friends if you found it interesting

English (auto-generated)

All

From the series

From Data with Zach

Machine learning

Computer programming

Presentations

Related

For you

Recently uploaded

Watched