

## WALK THROUGH

Creating Topics	→	Configuring Metadata	→	Configuring Topics	→	Managing Topics
-----------------	---	----------------------	---	--------------------	---	-----------------

What does the customer need to know?

ASSUME

- has AWS
- has QuickSight (conf)
- has knowledge of their industry
- mid-size company

building blocks!

Get Data

check to make sure everything is needed for the report / analyses

↓ → optimize for questions

Create Topic

↓

Label Datasets

→ metadata dictionary / DB

↓

(add additional data + / or filter settings?)

↓

Test out the topic  
w/ Q

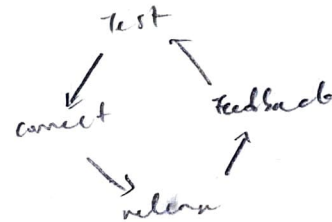
↓

Correct

↓

release

Feedback



Get Data

↓

Create Topic

↓

Label Datasets

↓

Test topic

↓

Correct

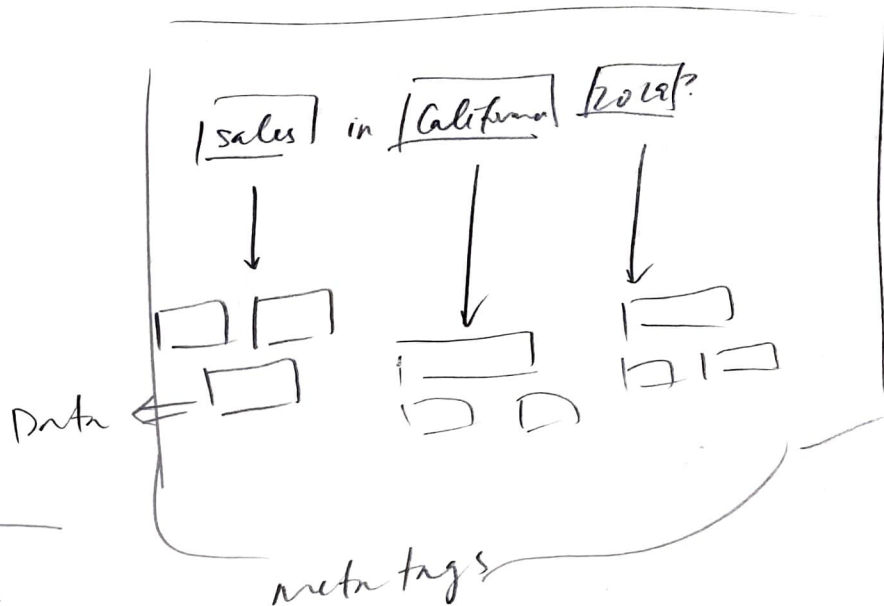
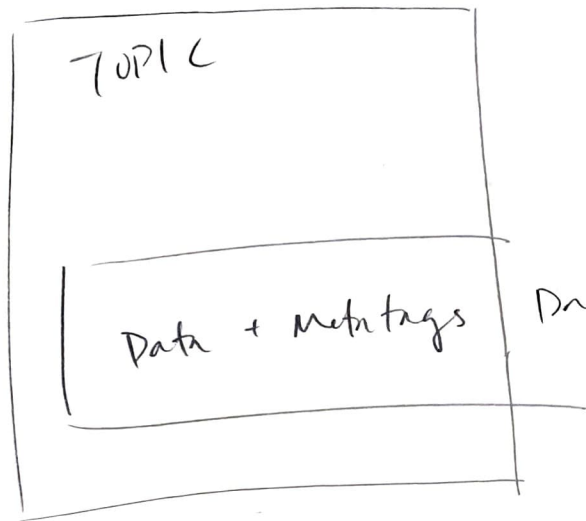
↓

Release

↓

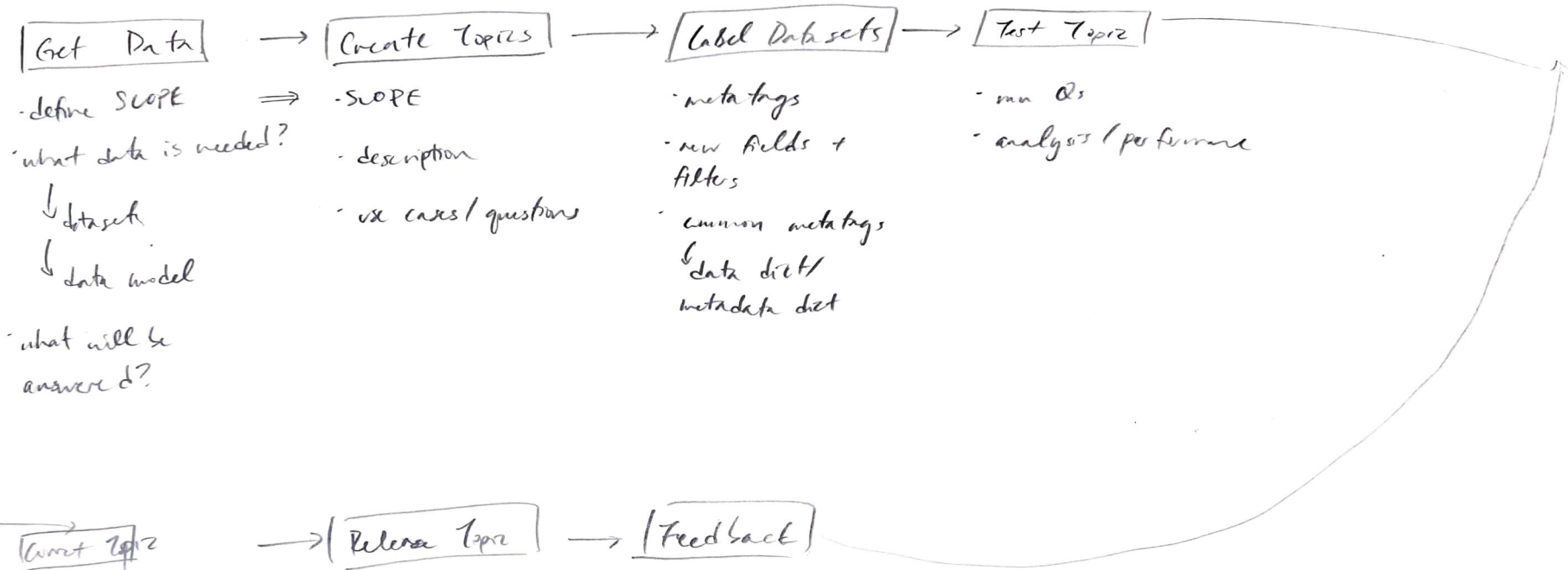
Feedback





Create Topic + Connect Data + Add Metatags + Test + Share + Feedback + Improve  
 ↓  
 considerations  
 + Review fields (are they req?)

# BUILDING BLOCKS



Pre reqs

DATA



TOPIC



LABELS



SQL CONFIGS



TESTING



CORRECTIONS



RELEASE



MANAGEMENT (?)

make your data  
SQL friendly

manage your  
topics

FEEDBACK



Users

A  
uid country

Subscribers

subid cost

Songs

sid song-name release-date genre num-streams duration cost

Streams

date time → date-time

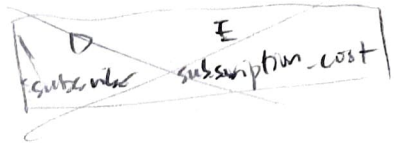
Artists

aid artist-name monthly-listeners

Ads

adid type info cost start-date end-date frequency

before subs down subs after subs  
new



uid → country, job, subscriber  
subscriber → subscription-cost

### Functional Dependencies

subscriber → subscription-cost  
(y/n)

$R_2(CDE)$   
 $D \rightarrow E$   
 $\{D\}^+ = DE$

$A \rightarrow BCD$   
 $D \rightarrow E$

$R_3(ABC)$   
 $A \rightarrow BC$   
 $\{A\}^+ = ABC$

$R_1(ABCDE)$

$\{A\}^+ = ABCD$

Subscriber  
subs\_id  
cost

cost

# Streaming Service Database

## Conceptual Model

can be queried with

0...\*

### Users

uid  
country  
join\_date  
subscriber  
(~~username?~~)  
subscription\_cost - new (\$15/month)

play 0...\*

### Songs

sid  
songname?  
release\_date  
genre  
num\_streams  
duration  
cost → def as cost per song stream

### Ads

ad\_id  
type  
information  
cost  
start\_date  
end\_date  
x frequency

calculate duration  
→ define as # of times aired?  
might be irrelevant

{ before - subs  
during - subs  
after - subs }

calc/new

↳ take out  
re calculate

### Artists

aid  
artistname?  
monthly listeners  
calculated new

streams associative  
entity  
date  
time

create

### Subscriber

0...1  
center

### Users

0...\*

play

### Songs

1...\*

create

### Artists

recurrent

Ads

1

# Logical Model

Users (uid, country, join-date, cost)   
 X old table → make new

Artists (aid, artist-name, monthly-followers)

Songs (sid, release-date, genre, total-streams, duration, cost, aid)

Ads (adid, information, type, cost, start-date, end-date)

Streams (uid, sid, date, time) → not subscribed

Subscribers (uid, cost) → NEW table

Recommendation (adid, uid) → not subscribed b/c not all users come slc of ads



Intro / Abstract

Previews

↳ data sets

Scanning  
↳ datasets  
↳ Data model map

Topic List

① Prep Datasets

② Create a topic

③ Develop ML first step

④ Test

⑤ Merge

⑥ Asking Questions +  
Getting Answers at  
Dashboard Q

① • Questions

• data model understanding

• finding data + if  
it's needed

• define terms + equations

② • questions to be answered

• create + description

• add datasets

• advice on naming convention

(all of them show up / updated  
or re-downloaded datasets come back  
as the same name)

• Time Bases

③ • config metadata

• config fields / values

↳ order to rank docs

↳ unique / dimensions

• Define TERMS

↳ i.e. popularity

④ • test topics w/  
g + diff methods

• link to agency

• disambiguation

⑤ • share

• report

• review

• actual vs. expected?

⑥ • asking questions  
↳ terms used

↓  
core correction

↳ question structure

↓  
disambiguation

Synonyms } Friendly Names } glossary?

Address Details

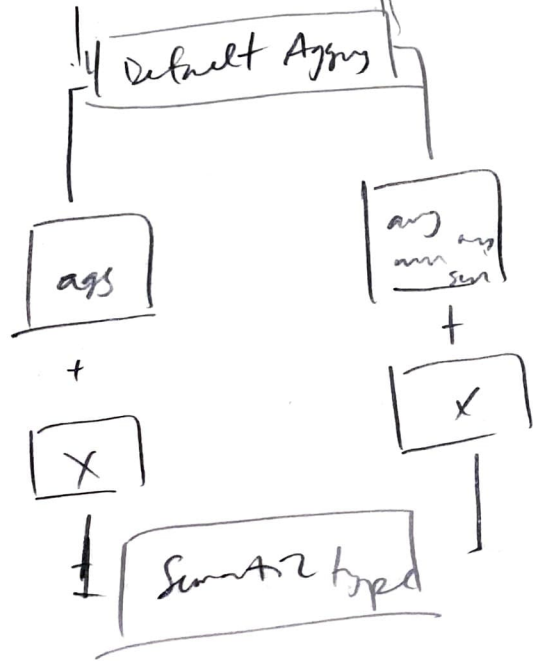
Description

Role



measure

Dimension



[vid] → use identification number



dimension (category)



count DISTINCT



no count ⇒ for which tables?

↳ users

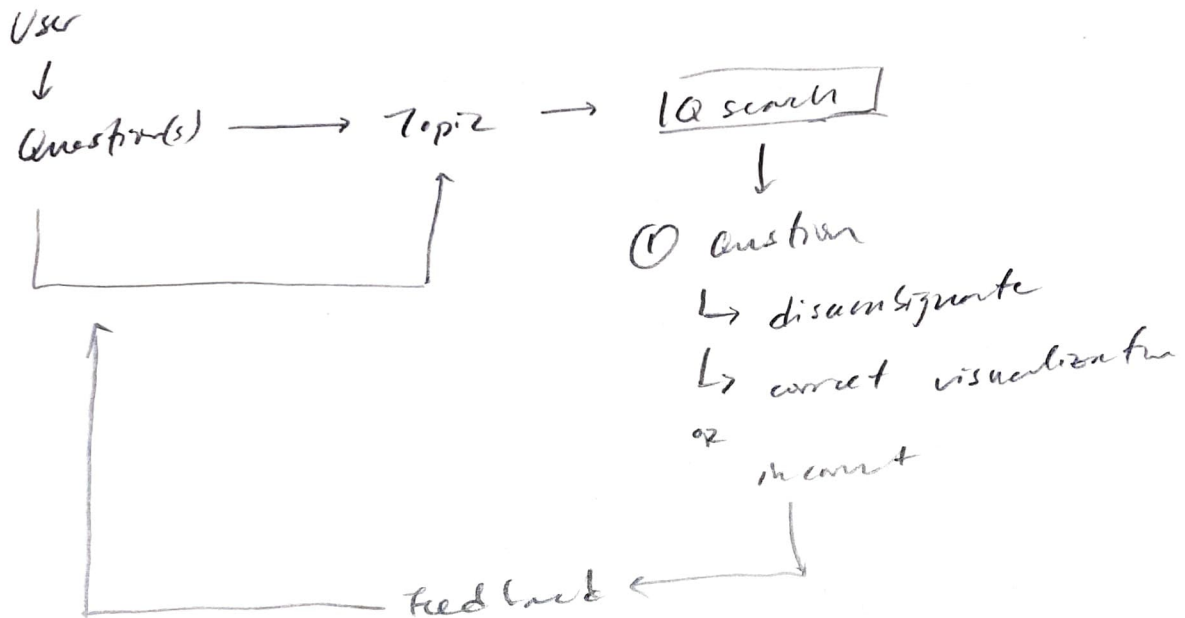
↳ subscribers

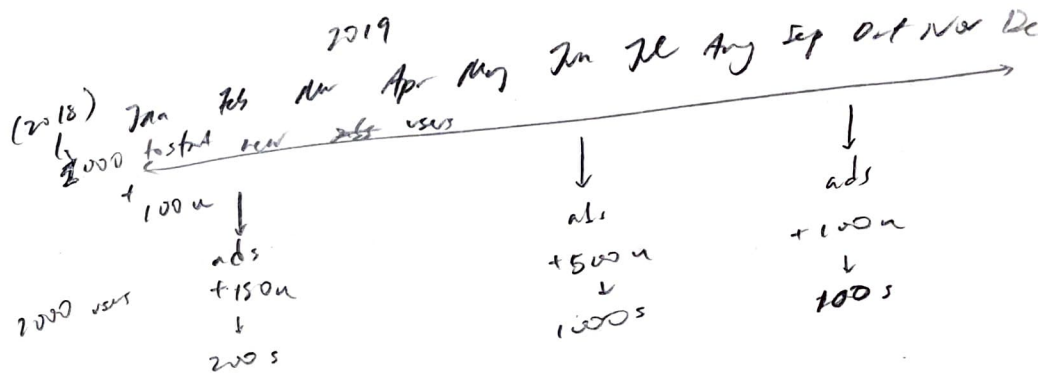
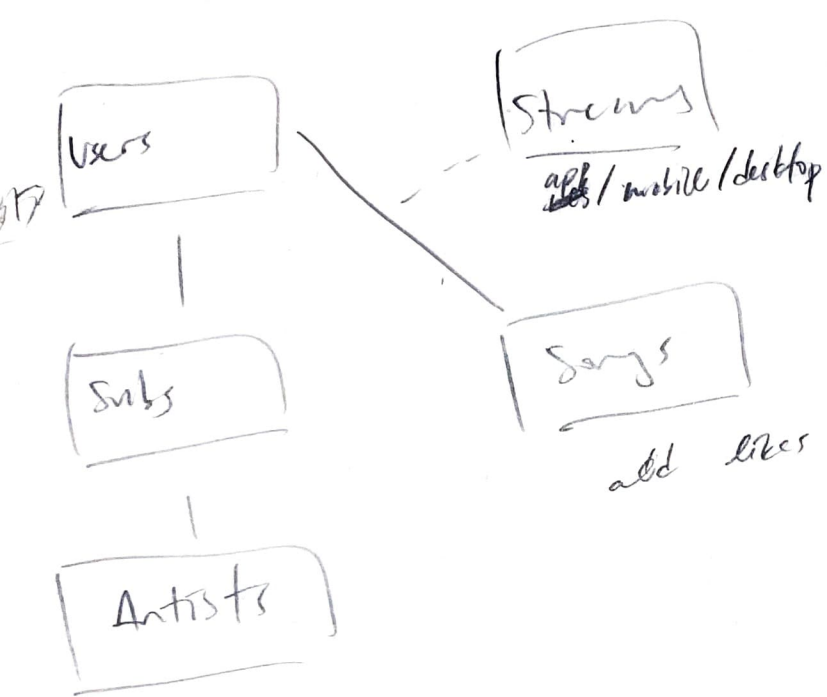
⇒ do count for:

↳ streams



semantic: identification





look at traffic

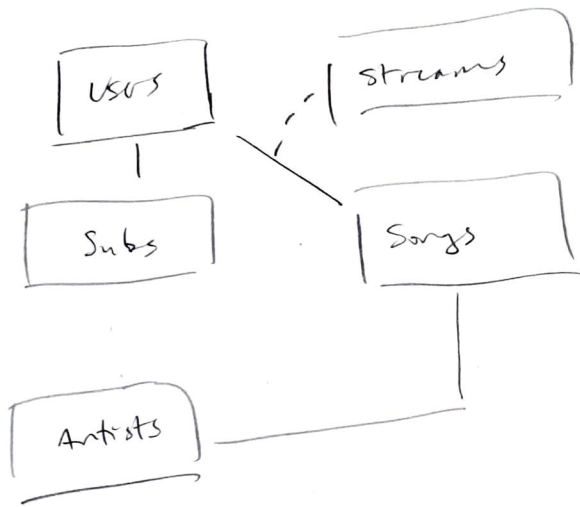
week / quarter

traffic numbers

Traffic	Profit
streams	set rate / artist (long term / short term)
likes	+ royalty / stream
	+ ads
	+ subscriptions

Data:

- 2000 - start 2000 users (2018) (thru Dec 31, 2018)
- 2800 - Jan 1 - 7: +500u + 500su
- 2750 - Feb 14: +150u + 100s
- 3550 - Jun 10-20: +800u + 1000s
- 3950 - Oct 20-31: +1400u + 250s



Campanas → X

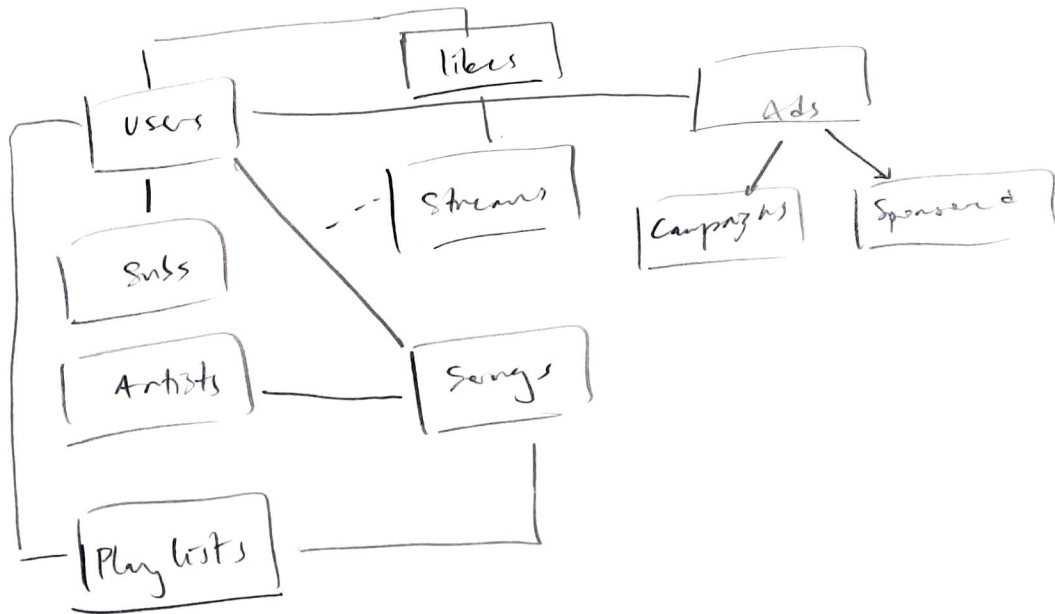
Users (uid, country, join-date, dob, age) x username / name →  
 x payment method →

Subs (sid, cost, ~~uid~~) start-date)

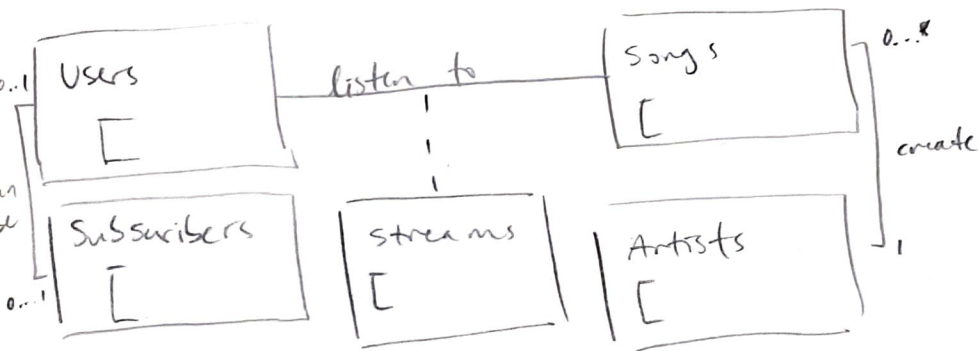
Artists (aid, name, monthly-fee, monthly-payment

Songs (sid, release-date, genre, title, duration, cost-per-stream, royalties, aid)

Streams (date-time, platform, ~~uid~~, ~~sid~~)



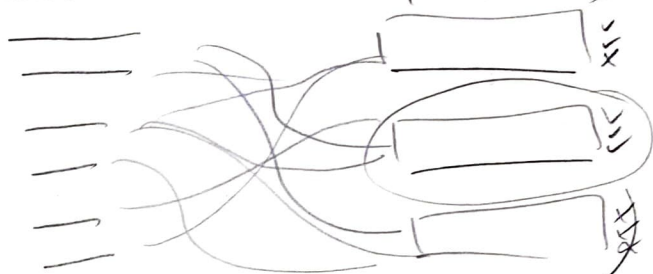
Graphs



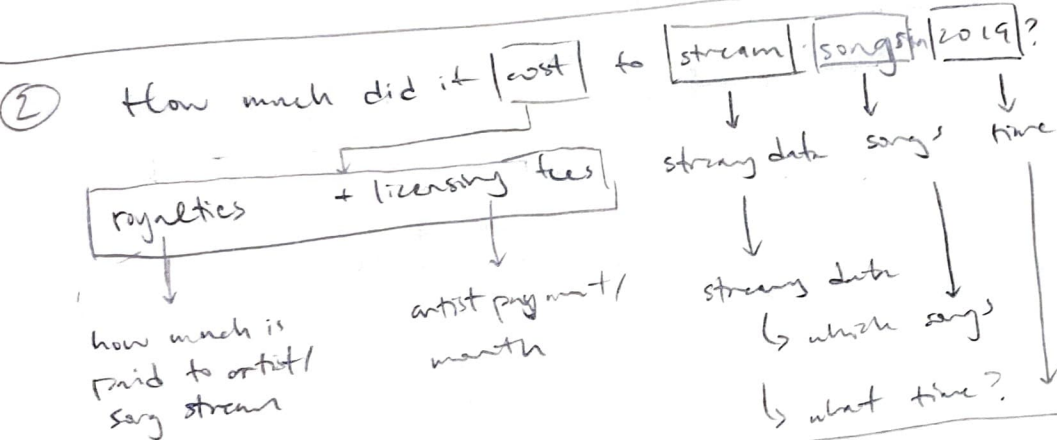
5) very Questions → Topic / Subject

Questions

→ quiz (practice)

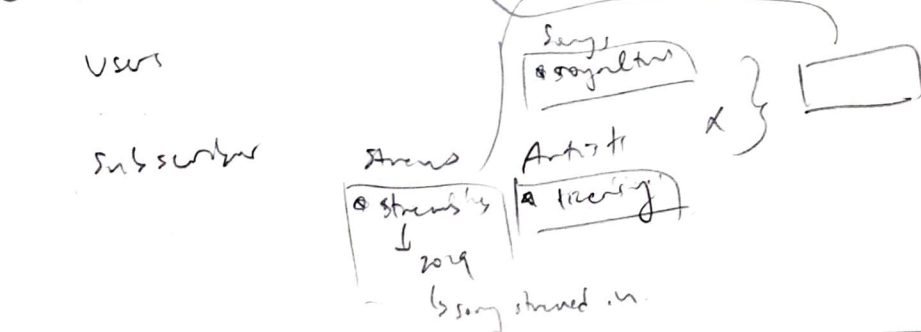


✓ comes more than 1 q  
✓  
✓



3) clean / uneven datasets - already in

4) How much did it cost to stream songs in 2019?



6) Dataset Include / Exclude

vid - identifier ✓  
vid - only if different table

vid - id  
vid - name X  
vid - signature X

change names! (or add date)

7) Dataset / Multiple / All

Dataset



# Graphics

⑧ Details needed/notes

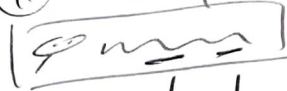
Measure



Dimension



⑪ Test to pass - summary



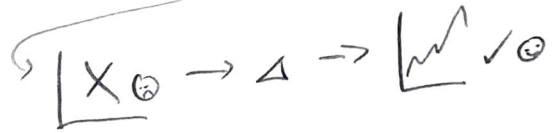
clarity

⑨ All details / diagram  
viz. summary dot & clarity

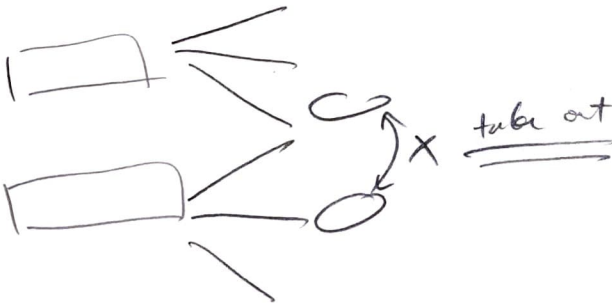


⑫ History Questions → Process

Question →



⑩ Synonyms





date Diff ( , , [MM]) \* subs\_cost  
start\_date      12-3

↳ assumes that subs cost doesn't change  
at least for that subs\_id

\* change date format for analysis

named entity

↳ filtering all fields

separate sub-type  
USD      Cnty  
↓      ↓

Subscription revenue 2019

(12-month subs) \* (subscription cost) ⇒ 2019

at time last @ [2019]  
Calculations

(12-month of 2019) & monthly sub. cost


• subscriber profit (2019)

• popularity — song — # of songs played + total streams field

platform — # of platform accesses in 2019

• traffic — platform —

date/time } streams

song — # of accesses on site → 

• top # / bot #

• rev by quarter

↓  
[rev]

how to identify?

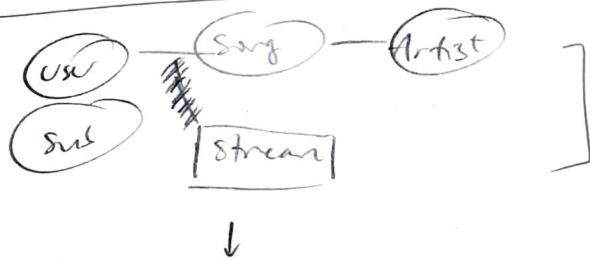
• most streams

• who streamed the most

• new users sing — 2019

• stream traffic in location

# Data Model Considerations



minimalistic, explain fields

data distribution + method of creation

Topic: Traffic Streaming / Revenue

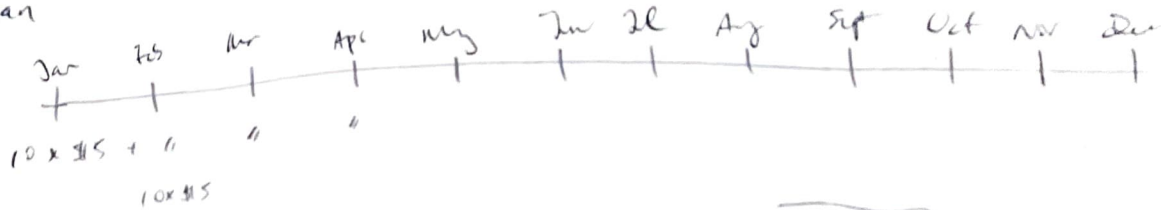
## Docs Considerations

- steps included  $\Rightarrow$  quite starting to a @ focus on sharing the product
  - $\hookrightarrow$  why / how to make different natural language friendly
- abstraction  $\rightarrow$  concept  $\Rightarrow$  task
  - $\downarrow$  dataset
  - $\downarrow$  index to main docs

10/10/10

user id	sub-cost

Jan



$$\boxed{\text{Jan \$}} + \boxed{\text{Feb \$}} + \dots$$

# of users that subscribed [x] month

+ # of users already subscribed (x-1, x-2, ...) months ago  
(after Jan)

OR

# of all subscribed users that month  
↳ use user ids? (in order)  
(not robust)

individual user



if this Jan

if

→ drill up / down to next / later  
→ can't do it across dataset?

- call field  
↳ select a dataset

- time basis

- finally name will show up as  
option if nothing is set

