Please turn [this video](https://m.facebook.com/Engineering/videos/data-quality-at-uber-how-to-get-data-right-at-uber-scale/1960560874116690/) into a ~1500 word technical blog post

It should include as much detail from the video as possible and follow the topics in the video as much as possible, so for example for the first section of the talk, it should cover:

* History of Uber’s data infrastructure
  + How they reduced the number of data pipelines and datasets
* Why they needed to built data quality/data observability platforms at Uber
* Uber’s Principles for data
  + Data as code (give definition)
  + Data is owned (give definition)
  + Data quality is known (give definition) and details from talk

- data observability at 2021

- tiered concept

Etc. etc. etc.

Here is a (rough) transcript of the video:

Thanks so much for arranging this and we had excited to be part of this summit and shared the learnings and the platforms that people that will put. And let me try to see what I have first. But it works and see her shot and you know I like the data quality and that the platforms for a while that will work and going to be working on open material and that I want to start off kind of how did the I the situation of you know miss it in the same data quality and that I can't read over. But we'll look into the legion is that of transportation but connecting the public and set likes connecting writers I was in the sun's each order what not you know behind all of this transportation platform is big data and the data sense and we built a really interesting services that so this particular picture kind of you know things that what we do it at a high level. So I'm going to go through that and you know go to the next steps on my people that was in that absolutely possible. So if you look at the the left side you have a key value to this which is nothing but a shot of my sequel. We have it all up on mobile apps and back in services and said that that's on the net this from there we build an ingestion framework which takes the concept saw that the lake and Lance online to be data to CDC into that lake which is poured by high. So when this process finishes it runs the political as it's called what it was so it's about it that is getting captured from on and it was then we have that I mean hosting which takes the lot it was and those more you know that mission they will send fact it was kind of a typical across the industry what happens and on top of that we build really nice you could use for the data science and engineering teams and officials to utilize the state. So the good thing about some service teams is so 7 to 6 in which it just a lot of interesting team it gives not a poet to our users. But also though it also generates a lot of data for us so as the business to end up you know go in going on that lake as well by the use of it and it's actually cost a lot of issues and a height of this that from we have 300 K that assets and the issues are that a good indication because no one knew what it existed and even if they can discover the data. You don't know who to talk to and understand the data so they end up in a creating their own say no question of the key and looking like so doing this. Yes, we consider mostly on let's say no scaling that inside so but kind of lacked investment to was that a product itself. So as I said like that of the patients that are just going to become a problem and so is that it won't be so nobody knew that it is kind of landing in the stables even if they didn't land and it's getting forward by dashboards which are not only used by some of that us and these use those expedients and also in the next on what is to understand what's going to go and how to better include these products we couldn't tell like no if the quality of the data that is landing there is great. So from there we apply some of the first principles. Do I read a bit of data culture. So as I said like at the height of the data platform we have it in the data sets by taking kind of a fundamental approach to and looking at how to it and the data at Uber with it just quite a bit of that from under supply plants with the different pipelines from seeing that a data set which are not owned nobody in the new who is the owner of those assets to and then that the data sets which are owned and we have been you know it on and give you can't use those and it on the 4 of those and make tricks that are pulling the state itself. So that we're going to what are the data for us Prince that as a court. What that exactly means is you know when that is getting created and dedicated on being changed it has to go to set in process. So still a change to detections and everything's on plate but it's going to not just can only at this problem off it happens. Right. You have a producer who created a will and the change the name of the column in the type of the column. Then that's a backward income for the change and the land and production and your pipelines are broken by then you get a similar mission but the problem is on the in what happened and you need to go back and in the sun of the june else to fix that. So by treating that as a code that that the facts has to be probably reviewed so in this too much change that we do in production go to review process. This review process not only thanks to the producers of the data but also anyone who's consumed that it so it's that's part of the data as a code. Now that is owned similar to what services when you actually building a service that city. For example, you're putting it in service in production it is on the team. A group of people that make shows that that you can go some is up and run similar living at getting it that a table it should have it on an association. The team should own that need to clearly define what is the what is the intent of the data but it's a lot of fact itself and only the way getting decent on the data. And finally that quality is known so that it will be here it's not just a what and the money that shuns this itself but what all this is it on the the fact you as the ownership of what what is it so they often that I think that you created and put this list not just for the data quality but also for the citizen on you know any books being final you're that fact and any instance being final on your that and how soon you can pick up by that and who is responsible for that. So from it's such a platform there and that has been well this know that some selling tools anyone get going there on the day that they don't dash was in their own assets. We're much more regulated and responsible that that from the people and we can transform the data comes from the cool. So these are kind of a foundation of principle that we started working on and forcing it. So we next night. It among the ticket and a sense and you cannot attend to get a sense you cannot and assume all of that that is important. Right so similar to we services. It will well I think the same to see that you can observe it is something from us to your concept. Example of it will burn to take a cup as a service here as seal because it is extremely important business need that will become something got got in fact not just the business but also use of experience. Similarly we have other services high even other things which are not so extremely important but doesn't it by using experience that much so we call the best year one. Right so we have been in the data on the facts as well you know that that that you have kind of building and promoting as a service for the rest of the company has an important story like is this that said he's being used to make patients on the business is this used to make in a production. But the conditions. In one morning some of them depending on the status. So if you want is an extremely important asset. And if I for example is an individually owned that generating a temporary staging tables which will you know doesn't have any guarantees at all. So we brought in the feeding concept we get pretty much all the data assets that over including not just the tables pipelines and models and dashboards as well. So with that we came from 10 that they will still be and here one day it was that it seemed important so then you know what is extremely important if your business you can actually focus on important data assets and not just the focus but you can run the pretty slow that I look at the resource for those tables so and so. And this could be another problem and of course the duplication at over as well. So we have fundamentally been about this that to providing not just a bit of that lot but that the session with this. And it's another problem that I kind of want to talk about in this case is that I know I know cooling that we had before and after the issues so we know we actually look at this that a problem to each individual team that over and a bit of that also look to know someone has a disability problem that the that love someone has a quality problem. And I'm only detection the only bill those specific use cases and what end up happening is made up and up the way that it in each of these tools because and that it is a political and in the up the student then the put them in the data itself that is no centralized. Once the store for make it so so we have this in this problem we build what we call us you make it up. It's only show these for all the metadata related activities so we build colors and build a meeting time to you make it on top of that we enable all these use cases. So if you go to the book until it talks about can plug in discovery but you find all the important information to us in the data that about this kind of the U.S. and top of you make it so you can actually go to pick what is your quality signals was that the year who won this. But the data assets lineage and you know and that and that but itself that enabled by the human to become and that it's and the next part is the nation. As previous because spoke images extending but not just for enough. You're understanding of how the data flows of the going and smart. But the image is extending put in the data quality enforcement as well should talk about in the next few lights. And the final part of the equation is one of the guarantees that they're giving out of the data but especially given to get us that's which we said like next to him but this is so we need we need that asset gets on to give one here to get us that includes to follow for that. But when they qualified for given to do we automatically on both in them into our data quality and in for certain checks and so the formation principle if you're given to your that that should be documented you should be owned this should be a page of duty. Uncle should be associated with the given and get to that assets and any others in the instance with automatically. Well in you know to go and other to us the team that owns so if you think of that is essentially a service at that point when you're like getting the data assets. So those are going to pay high level will be off and what's that I do we're like one of the problems that can come back and how did he can fundamentally a list of that equality of the data ownership itself. So in the next few slides and it will go through more details with the data quality itself and in the part that should be considered that with that Sunday you would think well. So it was on you. Yeah so as I was saying right. The 2 fundamental illness for the authorities that the one in spot and then the spot of the table which is the metadata catalog the lineage right. And it equality is kind of been done up of these 2. So I mean the most important thing is the dating but we started getting all the data sets to find out how many of these want to get it as it's important and as you can see the number of data sets where at the beginning of the 2020 didn't have a lot of data sets there and then now almost all the data sets appear and it turns out that of the one that he did us it we had only around 3000 of them are the one and the other. So those are the ones that are going to impact the business so we could focus more on them. So that that's a smaller number to work with. It's also going but it's it kind of separate of the noise there. So once we have these things right and that's just it. The set of the governance right and then we've been behind all the table owners to make sure they document the data sets that is this fiction. They could be at the owners the owner should not be due to a situation the teams and then they have the ability to be set up right so people know who to talk to one of the biggest problem was they are working on this data set and something has changed and they don't know who to reach out to because a lot of people. You know what the undisclosed you know fine right so we make sure he's on the data set and that is a big deal to reach out to them and as you can see in this light there is also freshness duplicates and cross DC consistency checks and come to that later so in the next slide. So then we had a system first of which we call trust OK where people were able to come and define their own checks right. So what happened was teams that had really good monitoring practices they had a lot of checks some people do not have checks at our level we were not able to answer the question is this check enough is this check not enough which tables are covered which are not covered which are so that was a big problem for us so we went we had a lot of conversations around that and we decided on four major categories of checks right and one is freshness and and in some places we call this completeness and the idea here is let's say you had a trip that happened at 12 p.m. right and then it flows through a Kafka topic and then it's ingested into the raw data table at 1 p.m. and then there are there is a model table and that is getting calculated at 3 p.m. so your freshness so when the table runs when the pipeline runs at 3 p.m. and then the when the pipeline runs at 3 p.m. the freshness is actually three others right at a high level that's how we measure the freshness the latency for the data to travel to the consumption point and that was a very important metric because the some teams might have it offered 24 hour freshness others might have offer 48 of freshness but when you are going as a downstream consumer right when when you have a dashboard or when you have a ML algorithm relying on this data you should know how fresh the data is right and then it's up to the owners to set their SLA but what we do is whatever SLA is being set by the owner right we hold them accountable by the by making sure we constantly monitoring these freshness right the second part is completeness. Completeness is kind of you know tied with freshness so it's not just how soon the data is coming is the data complete right you know you process the partition for yesterday but the partition for day before yesterday's incomplete right that there was an incident the pipeline broke and the data is still missing right so completeness gives make sure that all the data that is present in upstream is present in downstream. Cross DC Uber operates in two major physical regions and all the data is is backed up in the other region for high availability and durability right and the customers they don't always consume the data from one region they are kind of distributed across the regions so we want to make sure that the data is consistent at the physical level on both these locations so that as a consumer of this data right you should not be worried about the physical location of the data right you are free to consume it anywhere and duplicates is a lot of data we assume the records only exist once right a simple example would be user records that is only one user account so and a lot of pipelines kind of assume that and they have it so in my databases like my SQL right once you have a primary key you can't even insert a duplicate record but but with big data right that's not possible to prevent before insertion so what we do is after insertion keep checking for duplicates and it's based on the primary keys that is defined in the metadata catalog right now these are the basic checks that we define but on top of it we also allow users to write their own checks the own checks can be a custom query checking a particular thing because the fifth pillar that is not that is not mentioned is the semantic completeness of the data so it's the owners who know data database the owners as well as the consumer so maybe you're looking at certain distribution of a column it's an integer column right and let's say it's the trip amount right it doesn't make sense to be a negative value or it doesn't make sense to cost millions of dollars right so you could have a check on the distribution of the trip amount and it could be specific to a certain region right so we allow users to define their own checks something that we cannot enforce at an art level for all the tables right because the context is really important here and then they choose their schedule and they can go to the same data book that and they can define those checks right once all these checks are on board and right they show up along with the table so anybody who's coming to data book searching for a data right the same place that gives them the ownership the same place that tells them the description of the data is the same place where the quality of the data is also visible so they can look at all these checks they can pick a time range that they are interested in they can dig in and see how these checks are doing right and and it's also propagated to all the downstream systems so we have our own ML platform right so when somebody is training their ML model along with the ML quality report the data quality report is also attached or if it's a dashboard or it could be an ad hoc exploratory query right so you're a data scientist you want to build some new insights right so you're just going and running some phase and taking the data so in the query builder tool that we have we surface the data quality hey you're taking this table and you're looking at these partitions but during this partition dates there are some data quality issues right and when you go and search when we show the summary of all the tables we show like a green to indicate healthy right or red to indicate the data quality is not fine right so and what is the impact right so we defined downpins based on these checks so when the checks are met the data is good when the checks are not met the time period is how the downtime is calculated similar to what the previous talk the Monte Carlo they were saying so become of the SLA and as you can see once we started our governance once we started the data quality right you can see our SLA going up we are almost at double-nines now I would not start is to be triple nines in terms of tier one and tier two SLA's we're not dead yet but we kind of progressing there right so what is the future for us right so we want to enable more properties for the semantic test multi-dimensional test we also want to enable code reviews for tests when when users create the test we want consumers can use the data quality platform as a SLA negotiations platform so the consumers can go and create the test the owners can review it and approve it and then and the tests the SLA the thresholds of the test will page the owners right if it is not being met and there is a and we are also building features like anomaly detection which is pretty standard in the industry and expose that as another offering as well one final thing is something that Harsha is working on into open metadata I would like Harsha to do that. Thanks Ajay so as we talked about all these issues that it's not unique to the Uber itself these are the issues that are physical industry so with that you know to make these learnings and kind of solutions that we worked on Uber and previous companies you know we started a openmdata.org as a project so you know if you're interested in solving these issues want to take a look please come and take a look at open source project there thanks maybe