Audio file

2020-06-13_Training_Dkan_[Name redacted].m4a

Transcript

00:03:43 Speaker 1

There we go.

00:03:45 Speaker 2

Yeah, later we can edit these and.

00:03:47 Speaker 2

Make them available.

00:03:48 Speaker 2

To folks done anything with.

00:03:49 Speaker 1

I think.

00:03:50 Speaker 2

Them yet but.

00:03:51 Speaker 3

[Name redacted], I think if you select spotlight.

00:03:53 Speaker 3

It'll only show you.

00:03:56 Speaker 3

But I know.

00:03:58 Speaker 3

Some people, I think, turned off their cameras but might have them on otherwise.

00:04:01 Speaker 3

But whatever.

00:04:03 Speaker 2

Alright, let's see training breakout. I don't.

00:04:05 Speaker 2

See spotlight in the options.

00:04:11 Speaker 1

Back up.

00:04:15 Speaker 1

Hey, come on. Let go of the thing.

00:04:20 Speaker 2

Stop sharing. There we go. Let me start sharing.

00:04:26 Speaker 4

OK, cool.

00:04:28 Speaker 2

So welcome to today's training disinformation artifact, decan and hue.

00:04:33 Speaker 2

We're going to be talking about.

00:04:36 Speaker 2

What sort of?

00:04:37 Speaker 2

Things we do with dcan what it is.

00:04:39 Speaker 2

And how to use it?

00:04:41 Speaker 2

So first of all, what is?

00:04:42 Speaker 2

Deccan Well, if you remember back to the early part of the 21st century, there's a big push to put publicly available data that was collected by governments of various shapes and sizes.

00:05:00 Speaker 2

On the Internet.

00:05:01 Speaker 2

So that citizens could see it and do.

00:05:05 Speaker 2

Interesting things with it and be better informed participants in their democracies and this.

00:05:12 Speaker 2

Was a push that resulted in data.gov. This is a push that resulted in a number of cities and municipalities and state governments and other countries putting their data on the Internet for everyone.

00:05:26 Speaker 2

To see and.

00:05:27 Speaker 2

One of the early platforms for doing that was something called SECANT, which stands for.

00:05:32 Speaker 2

Comprehensive Knowledge archive network.

00:05:35 Speaker 2

And it's great, it's Python, it's Python.

00:05:39 Speaker 2

Web application stack built on Django, which is great if all you want to.

00:05:43 Speaker 2

Do is share open data?

00:05:45 Speaker 2

But there were a number of state governments and.

00:05:50 Speaker 2

Number of federal.

00:05:50 Speaker 2

Government agencies who were using a platform called Drupal, which is a PHP content management framework.

00:05:57 Speaker 2

And they did not want to have to reimplement an entire new web application stack, so a group of Drupal developers put together a feature implementation of feature. Complete implementation of Ccan and Drupal and called it D.

00:06:13 Speaker 2

And so now we have this platform being used by the United Nations, by the federal government, help the Human Services Agency, various state and local governments, a number of foreign countries are using this platform called Deccan.

00:06:29 Speaker 2

We're using it.

00:06:31 Speaker 2

Because it has the ability to share multiple data types like CSV, JSON, PDF, ZIP files, raw images, you name it. We can do data visualizations on CSV and JSON using the recline JavaScript library, so you can do charts and graphs with.

00:06:49 Speaker 2

Geo location data. So if you have latitude and longitude coordinates as part of Geojson in in a JSON upload you can do map visualizations and geolocation visualizations. So you can do things like choropleth maps where you can show heat mappings of various.

00:07:09 Speaker 2

Data concentrations. You can do things like show you know where different sites of different events are happening, right? And the other reason we're using it is because it has a fully rest compatible API.

00:07:23 Speaker 2

We want to use it as a data warehouse specifically for TLP white data. At the moment it allows us to store incident artifacts until we have time to process and analyze them, which is going to be important going forward as we deal with COVID-19 disinformation. Sometimes stuff gets put out very fast.

00:07:43 Speaker 2

Very hot and heavy and we don't have time to catalog it in the moment. We do want to capture that data so we can go back and do more in depth analysis on it later in time and then identify things like patterns.

00:07:56 Speaker 2

The actors behind various campaigns.

00:08:00 Speaker 2

Patterns and language patterns and posts etc.

00:08:03 Speaker 2

It's also good for storing amorphous stuff that doesn't really fit anywhere else, right? So if we have a bunch of images that we want to throw up into a data set in Deccan because they don't fit well into Google Drive or GitHub, that is something we are more than empowered to do.

00:08:22 Speaker 2

The biggest benefit to this, and what we originally decided to do when we started COGS that collaborative back in January is we wanted to set this up to make it easy to share data sets with other data scientists. And this info response teams. So using this model, if we find something that's not necessarily in our wheelhouse. So let's say we find more protest.

00:08:42 Speaker 2

Info stuff we can.

00:08:43 Speaker 2

Throw it in Deccan and then.

00:08:46 Speaker 2

Alert teams that are working on protest disinfo stuff and say here's this data that we found. Here's a link to it. Go grab it and do what you need to do with it. And that's true for other things as well, right? So anti VAX, which is, you know, tangentially related to COVID-19 stuff related to the election and basically lets us grab data artifacts related to disinfo campaigns, throw them someplace where other people can access it.

00:09:09 Speaker 2

And then share that with them.

00:09:12 Speaker 2

Now we could be using this for more stuff than we currently are. Like I mentioned, we're using it to store top white data. We could do more visualizations with that data, so if we for example get a bunch of information on hashtags, we could in theory using recline do.

00:09:29 Speaker 2

Hashtag clouds and word clouds so we can see which hashtags are more prevalent.

00:09:35 Speaker 2

With some permissions modifications, we could be using it to store top green or amber data, right? So stuff that we want to share within CTI but not publicly or stuff that we want to share, say with Reddit or Facebook or Twitter, right, but not publicly available. So we could modify the permission schema for Deccan to store that data.

00:09:57 Speaker 2

And then due to the API, we could also be using it as a slack bot endpoint so we could be setting up Slack bot to automatically throw stuff into Deccan after it gets added to the Internet archive. Or we could be using it to tie data sets in Deccan to incidents in the hive or in Miss, right? So we have a number of options here which haven't.

00:10:15 Speaker 2

Really capitalized on it yet?

00:10:17 Speaker 2

Limited time and people power.

00:10:21 Speaker 2

So any questions before we go into demonstrations?

00:10:28 Speaker 3

It's just a tool that you guys used prior to starting COGS tech collab, and if so, kind of how did.

00:10:34 Speaker 3

You guys use it so.

00:10:35 Speaker 2

In the interest of Full disclosure, I used to work for the company that developed.

00:10:42 Speaker 2

And before I left to go to tenable.

00:10:45 Speaker 2

So some examples of how it has been used in the past. One of my favorite examples is.

00:10:53 Speaker 2

The the city of San Antonio, TX.

00:10:57 Speaker 2

Put all of their civic data on dcan they they put up an open data instance, they dumped a whole bunch.

00:11:04 Speaker 2

Of data into it and a group of cyclists.

00:11:07 Speaker 2

Basically went and.

00:11:08 Speaker 2

Pulled two data sets. They pulled the geographic.

00:11:13 Speaker 2

The latitude longitude data for every bike path and bike lane in the city of San Antonio and then.

00:11:20 Speaker 2

They went and pulled.

00:11:21 Speaker 2

The complete record of traffic collisions and focused on bicycle vehicle collision specifically and they were able to identify the 10.

00:11:33 Speaker 2

Most incident prone and deadly intersections for bicyclists in the city, and they went to a City Council meeting and said, look, these are the 10 most deadly intersections for cyclists in the city of San Antonio. If we had $100,000 for improving, you know, visual visual markers that you're approaching.

00:11:54 Speaker 2

Bike, bike, track, bike path intersection and you know, improving visualizations and improving.

00:12:00 Speaker 2

Like putting lights.

00:12:01 Speaker 2

Up for when people are crossing, we think that we could reduce fatality significantly and they were actually able to reduce fatalities by something like 70%.

00:12:11 Speaker 2

Just by looking at civically available data.

00:12:15 Speaker 2

Another example, the city of Minneapolis is able to put together a adopt A fire hydrant program about six years ago during a huge spate of blizzards in the city where they were running into this problem where they could not.

00:12:29 Speaker 2

They literally could not find fire hydrants because there was so much snow on the ground and they were able to during a hackathon, build a adopt A hydrant website taking the the geographic location data for every fire hydrant in the city and putting it on a website where people would adopt A hydrant near their house.

00:12:48 Speaker 2

And then when it snowed, they'd get a text message notification saying, you know, it's snowing. Please clear off the area around your fire hydrant. And I think they saved something like 100 or 200 extra houses from burning to the ground that winter due to not being able to fire, find a fire hydrant.

00:13:03 Speaker 5

Yeah, there's there's, like, a whole civic data pedigree that underlies.

00:13:07 Speaker 2

Yeah, but the neat thing about this platform is because we can put stuff.

00:13:08 Speaker 5

The development of this right?

00:13:15 Speaker 2

Into it and have it be web accessible. There's not as big a barrier to entry for people to to find it and go digging and playing with data. It's primarily geared toward the data scientists that we're, you know, training and trying to bring in so that they can go in and pull out data, download it, run it through Python, do whatever or, you know, set up visualizations.

00:13:35 Speaker 2

Directly and decan and then.

00:13:37 Speaker 2

You know, play with it there.

00:13:40 Speaker 2

We just haven't capitalized on it very much yet. That's starting to change.

00:13:44 Speaker 2

We've got a.

00:13:45 Speaker 2

Couple of folks on the CI bot team who are starting to play with the Deccan API so that we can do things like.

00:13:50 Speaker 2

Automatically put data in.

00:13:53 Speaker 2

Any question for you, [Name Redacted]?

00:13:54 Speaker 3

Yeah, that's awesome. It also makes me wonder, and this isn't necessarily a question, but more just a thought. You know, as [Name redacted] and other folks are talking about doing hackathons, if we have kind of a relevant incident or set of data in Deccan already, we could give folks access to that and ask them to play for it, play with it, especially if it's you know TLP [Name redacted] or.

00:14:15 Speaker 2

Right.

00:14:15 Speaker 5

Yeah. Data that's in there or also playing with getting data in like managing the front end of an incident and building the tools to get it into the system.

00:14:24 Speaker 1

Right.

00:14:28 Speaker 5

For response workflow pipeline analytics.

00:14:34 Speaker 3

Thanks so much, [Name redacted].

00:14:35 Speaker 6

No worries all.

00:14:36 Speaker 2

Right, so now I'm going to show you.

00:14:39 Speaker 2

What it actually looks like?

00:14:40 Speaker 2

So I'll do a quick tour through here before.

00:14:42 Speaker 2

I show you how to do things like sign up for an account.

00:14:45 Speaker 2

So if you go to data.conceptcollab.org.

00:14:48 Speaker 2

You end up.

00:14:49 Speaker 2

On the front page here and most of the magic occurs here in the data sets page, so it'll show you everything that we've currently got in here.

00:15:02 Speaker 2

There's a bunch of stuff, so we have incident paper Flyers. We have one on COVID 9 or COVID 5G stuff that was on Reddit. Stuff on pandemic. There's a thing that we started following back in.

00:15:15 Speaker 2

March related to a Stafford Act Stafford Act hoax that was going around.

00:15:24 Speaker 2

But one of the big ones that we're going to look at right now is some CS fees related to COVID-19.

00:15:32 Speaker 1

Some, I think there's [Name redacted] that put this in.

00:15:36 Speaker 2

Either [Name redacted] or [Name redacted] anyway.

00:15:40 Speaker 2

So you can see here that there are.

00:15:43 Speaker 2

For resources attached to this data set, so the terminology that is used with Deccan is a data set, is anything related to one type or specific congregation or aggregation?

00:15:59 Speaker 2

Of data, right? So.

00:16:00 Speaker 2

In a civic context, it would be stuff like.

00:16:04 Speaker 2

You know, intersections, police arrest records, traffic fatalities, property tax data, you know, vaccination rates, stuff like that. And then resources are individual pieces of data that are attached to a data set. So you can see here there are four resources attached to this data set on COVID-19.

00:16:31 Speaker 2

So if you have an account, you can do things like upload data, manipulate data, add data to data, or add resources to data sets. You can build visualizations today. I'm just going to focus on.

00:16:44 Speaker 4

How to set up an?

00:16:45 Speaker 2

Account how to add data?

00:16:49 Speaker 2

How to start a new data set and how?

00:16:50 Speaker 2

To add data to an existing data set.

00:16:54 Speaker 2

So for registering an account, what you do is you go to website, you click on register.

00:17:01 Speaker 2

And you fill in a user name and an e-mail address and you click create new account and what that does is it sends an e-mail notification to me that you have signed up for an account and it is pending activation.

00:17:14 Speaker 2

What I would like you to do if you sign up for an account is then ping me in slack. It gives me an extra layer of verification that it's not just some random person who found this website and wants an account. We want to try and keep the ability to add data to people who are either members of COGS at collab or CT I league.

00:17:35 Speaker 2

If you sign up username e-mail address, click create new account, I get an e-mail notification and as soon as you ping me and Slack I go in and activate your account and you'll get an e-mail with a one time link to create a password.

00:17:48 Speaker 5

Is there any touch with Jr.

00:17:51 Speaker 2

Not at the moment, although that is something I need to talk to her about. Basically to assume that once you're in the triage team, right? Yeah.

00:18:04 Speaker 2

Is a good question, but yeah.

00:18:07 Speaker 2

Also, because it's all TLP white.

00:18:08 Speaker 2

Data is not a lot.

00:18:11 Speaker 2

Yeah, if we modify.

00:18:13 Speaker 2

The permissions so that we can store green and amber data in here and then that's something.

00:18:17 Speaker 1

We're going to want to look into.

00:18:21 Speaker 2

So for now.

00:18:22

I'm going to.

00:18:22 Speaker 2

Go ahead and log in as me.

00:18:30 Speaker 2

So this is what it looks like when you log in.

00:18:32 Speaker 2

And show you all the.

00:18:33 Speaker 2

Data sets that you've created, anything that you've tagged.

00:18:36 Speaker 2

It shows you.

00:18:39 Speaker 2

How many data sets you have? All that fun stuff?

00:18:42 Speaker 2

And then there is a menu of stuff up here on the top.

00:18:48 Speaker 2

To add content, let's start with adding data sets. You go to add content and click data set.

00:18:58 Speaker 2

And then it's going to take you to the add data set.

00:19:00 Speaker 2

Page so.

00:19:04 Speaker 2

What I'm going to suggest is that you make your titles and descriptions as complete and informative as possible, because in the interests of documentation this is stuff that we're not going to touch until much later. And the better your.

00:19:17 Speaker 2

Documentation is in the moment.

00:19:19 Speaker 2

The happier you're going to be in three weeks when you have to go and figure out what exactly this thing is, right?

00:19:26 Speaker 1

So we'll call this an example saying data set.

00:19:30 Speaker 1

Data set.

00:19:32 Speaker 2

#2 right.

00:19:39 Speaker 1

Training add data to.

00:19:47 Speaker 2

You have the.

00:19:48 Speaker 2

Ability to format text with markdown. It's not.

00:19:50 Speaker 1

Going to be.

00:19:51 Speaker 2

That useful in terms of where you're throwing data, because if you're writing a longer document, chances are we're going to keep it in GitHub or in Google Drive. But you have that option.

00:20:02 Speaker 2

With regards to tags, it's helpful to tag stuff so that we can go back later and find more in depth categorizations and sort of sort more easily.

00:20:12 Speaker 2

So for example, I'm just going to take.

00:20:13 Speaker 2

This as 5G anti VAX.

00:20:17 Speaker 2

This info and then maybe if I need to add a new tag so if.

00:20:21 Speaker 2

I want to say something like.

00:20:24 Speaker 2

Stan, right. So let's say I'm tracking 5G anti VAX disinfo in Turkmenistan. I can create a new tag just by typing it in and hitting enter.

00:20:33 Speaker 2

OK, we're not really worrying about groups or topics at the moment. Those are just different ways of categorizing and sort of putting data into buckets.

00:20:43 Speaker 2

You are worried about license because it's TLP light TLP white. We want to make sure that it is adequately licensed for sharing with different groups. We are doing Creative Commons Attribution share like.

00:20:57 Speaker 2

Because then it gets to make sure that credit comes back to COG second CTI.

00:21:03 Speaker 2

Once you've got all that squared away.

00:21:05 Speaker 2

You click next add data.

00:21:08 Speaker 2

And that will take you to the add resource page. Now there's two types that we're primarily working with. There's three tags here, but we're really only using two of them in cognac and CTI.

00:21:20 Speaker 2

If the data you are adding to this data set is contained in one file, right? So let's say it's a CSV full of URLs or you've got a screencap image, or it's a zip file with the.

00:21:35 Speaker 2

Contents of a.

00:21:35 Speaker 2

Of a scrape of, say, like a Twitter account or a subreddit or what?

00:21:41 Speaker 2

Add that as an.

00:21:41 Speaker 2

Upload right so you just click browse or you can drag it straight in and I'll do that right now and see if.

00:21:47 Speaker 1

I can.

00:21:47 Speaker 1

Find a yeah, I.

00:21:49 Speaker 2

Was going to do the Mueller report.

00:21:51 Speaker 2

Right. So you have a 10 gig file limit.

00:21:56 Speaker 2

I would be amazed if anything that most people upload is that big, but I'm using them all. A report as an example because it's 150 megabyte PDF.

00:22:07 Speaker 2

And so that will give you a regular progress bar.

00:22:11 Speaker 4

If it's a CSV or.

00:22:11 Speaker 5

Since we're waiting, do you mind if I ask a quick question? Just going to say in terms of the data use and the attribution and all that sort of stuff, do you? There's a cognac has an entity and are you asking people participating for like a?

00:22:14 Speaker 2

Yeah, go ahead.

00:22:30 Speaker 5

Any sort of data use agreement like you might in an open standards standards.

00:22:36 Speaker 2

Yes, one of the things we do when people sign up and come into slack is we asked them to fill out a form saying that they read a license, anything under CC attribution share like.

00:22:47 Speaker 5

OK. And or their employer?

00:22:52 Speaker 5

Because sometimes, sometimes.

00:22:52 Speaker 2

That's a thing. That's a thing related to their employer. So if they have a specific non disclosure agreement that is going to be between them and.

00:23:02 Speaker 5

Their employers? Well, no, I mean like they're so sometimes certainly in the open standards world, right where I'm contributing code. And I work for [company redacted]. I might be contributing.

00:23:13 Speaker 5

Threat intelligence because and I'm with.

00:23:16 Speaker 5

Wolf, wolf eye. Let's call it.

00:23:18 Speaker 5

Right, yeah.

00:23:21 Speaker 5

We don't want open data.

00:23:24 Speaker 5

To be encumbered? Potentially.

00:23:27 Speaker 5

By a claim from so that so the in the agreement, they usually will also say that you know what I'm contributing is either my own or it's not got anything to do with my employer or it's cleared with my employer.

00:23:40 Speaker 5

Where my employer releases any claim as well, right? So that certainly an open standards, right when they're implementing a standard in open source they're doing, they're doing that clearance because they want to make sure that the the licensure is.

00:23:54 Speaker 5

Is not encumbered to anything.

00:23:57 Speaker 5

Like there's no poison pills.

00:24:00 Speaker 5

Right in in the the the chain of ownership really at the end of the day something.

00:24:04 Speaker 7

Basically, before you contribute to any of this, you want to make sure that your employers are OK with it and are aware that you're not crossing any boundaries with your employment.

00:24:05 Speaker 5

To look at.

00:24:13 Speaker 5

Employment. Yeah. And you may you may want to make sure that's in the agreement that they're so it's something to look.

00:24:18 Speaker 5

At anyways.

00:24:19 Speaker 2

With regards to any of the data that's being edited.

00:24:23 Speaker 2

99% of that is going to be stuff that has nothing to do with your employer, right? It's stuff that you're pulling from Reddit or Twitter or Facebook.

00:24:30 Speaker 5

Yeah, it might be open. It's open source generally, right? Yeah.

00:24:33 Speaker 2

It's in the public domain or it's?

00:24:35 Speaker 2

You know complies with the.

00:24:36 Speaker 5

But if I'm on the clock and my employer is paying me to find that, and it's open source and I'm putting it in your database, they still potentially have a claim. So that's what I'm suggesting is that you want to make sure that you're covered from that, so that a really good data source that you might wind up with is actually.

00:24:52 Speaker 5

OK, for you to licenses as Creative Commons and it's not some kind of encumbrance.

00:24:57 Speaker 2

Yeah, makes perfect sense.

00:25:00 Speaker 4

OK.

00:25:02 Speaker 5

Considering the volume, we might wind up generating.

00:25:04 Speaker 7

We can. We can look at adding that adding that to our terms of service and that's in my realm.

00:25:10 Speaker 2

Yeah, Glenn.

00:25:11 Speaker 3

Awesome. Thanks.

00:25:13 Speaker 2

So let's see here. So it's going to.

00:25:15 Speaker 2

Complete the upload.

00:25:17 Speaker 2

OK, it's a test, right?

00:25:20 Speaker 2

If it's a JSON or CSV file, you have the options to do some different visualization previews.

00:25:27 Speaker 2

Not going to worry about that right now. That's. That's our different training at a different time.

00:25:32 Speaker 2

If so, that that works for for points of data or pieces of data or incident artifacts that all fall into one file or all?

00:25:40 Speaker 2

Fit in one file.

00:25:42 Speaker 2

If it is an artifact that is, for example, something we threw in the Internet archive, so we found a tweet when we found something on Facebook, we had it indexed in the Internet Archive.

00:25:53 Speaker 2

Then you can throw the URL in here, right? And that'll that'll work as a. Basically, it'll just present a link to that artifact.

00:26:04 Speaker 2

Remote files is not something we really use. That is for for things like sharing or cross linking to things that got uploaded to data.gov as an example, or another municipal or.

00:26:17 Speaker 2

Civic open data site.

00:26:20 Speaker 2

Because of the nature of most of this stuff on social media being subject to being deleted.

00:26:30 Speaker 2

I'm not going to recommend that we use that either throughout the Internet Archive and then post the URL or grab a screencap or scrape it using a scraper tool, put it in a zip.

00:26:39 Speaker 2

File and upload it that way.

00:26:41 Speaker 1

OK.

00:26:42 Speaker 2

Same thing is true here for titles and descriptions. You don't have to be as detailed because it's being attached to the data set.

00:26:52 Speaker 2

What I would suggest though is maybe writing down stuff like so an example being.

00:27:04 Speaker 2

So describe what it is.

00:27:08 Speaker 2

TNC 11.

00:27:17 Speaker 2

Something like that, right? Just where are you grabbed it from? Ohh, maybe a brief description. If it's a zip file, it's like the ZIP file contains a bunch of artifacts related to a subreddit that I found detailing COVID-19 and anti VAX. You know, organizing.

00:27:33 Speaker 2

The more data we have on where you grabbed it from and what you know, like if it's a zip file, what's in it, or if it's a JSON file with the different you know keys for the the the objects or arrays in the JSON are the more.

00:27:51 Speaker 2

Useful it is to us.

00:27:54 Speaker 2

Don't have.

00:27:54 Speaker 2

To worry about format that will be automatically detected.

00:27:58 Speaker 2

And then data set, it will automatically attach it to the data set. If you are creating a new data set.

00:28:05 Speaker 2

I will show you how.

00:28:06 Speaker 2

To do that in just a second if.

00:28:08 Speaker 2

You're adding a.

00:28:10 Speaker 2

Resource to an existing data.

00:28:11 Speaker 2

Set if you want to add another one. If you've got multiple files you want to add, you just click save and add another. Otherwise you click.

00:28:19 Speaker 2

Don't worry about initial info for now. That tends to be more for things that have geographic indicators, stuff that has to be specifically licensed because it is, you know, like you were mentioning before.

00:28:37 Speaker 2

You can change the licensing between data sets and individual resources. For now, we're not going to worry about that.

00:28:46 Speaker 2

So you click save.

00:28:48 Speaker 2

OK. And this is what you see. So you see the what the resource specifically looks like shows the MIME type, file upload timestamp etcetera.

00:28:57 Speaker 2

If I go back and look at this data set, it'll show me that I have one resource attached to it. What the tags are, the UID, identifier, license, public access level, etcetera, etcetera, etcetera. And I also have links for posting this and sharing it. I don't really need to worry about.

00:29:13 Speaker 2

If I want to add an additional resource to this data set, let's go find it. Either go get it out of the data sets listing here, or if it's something that you've added to before, it'll show up when you log in.

00:29:27 Speaker 2

And then you click add resource.

00:29:31 Speaker 2

And the same procedures apply, right? So you upload your file or you drop your URL in, fill in your fields.

00:29:38 Speaker 2

And then you click save.

00:29:40 Speaker 2

Now one thing that I do want to point out is that you can add the same resource to multiple data sets.

00:29:50 Speaker 2

So if you find something that falls under election security and anti VAX as an example, you can have one data set. So you can say scrape A subreddit, it's got data on both, you throw it in as a resource and then you can attach it.

00:30:07 Speaker 2

To multiple data sets.

00:30:10 Speaker 2

So that you don't have to duplicate, it's not something we're really worried about right now. I think we've got like 100 gigs of data storage available.

00:30:16 Speaker 2

To us, for files that are usually 10 kilobytes or less, so we're not worried about it at the moment, but it is something that you can do if you don't.

00:30:26 Speaker 2

Want to have to duplicate resources.

00:30:32 Speaker 4

With regards to the API slides.

00:30:36 Speaker 1

And go back to.

00:30:39 Speaker 2

So the Deccan data Set API is how you programmatically interact with data in Deccan. The standard Restful API. The documentation for it is at Deccan dot. Read the.

00:30:50 Speaker 2

Docs dot IO.

00:30:53 Speaker 2

If you want to look at some example code for how to work with the API, you can go to the Cogset collab GitHub organization and then open the dcam bots repo and you'll see some code that I've put together.

00:31:06 Speaker 2

Or how to work with that API OneNote don't use production for this. If you want to play with the API and build a slack bot or build a tool for adding lots of data in one go, use the dev site. I can easily move the OR.

00:31:27 Speaker 2

Populate the development database with logging data from production. So just let me know. I'll set that up for you and then you can play with data to your hearts content. The other option is you can use. There is a set of tools related to Deccan that was put together by the Deccan team which will let you set up a fully functional Deccan.

00:31:43 Speaker 2

Instance in Docker so that you can just basically spin that up, play with it.

00:31:47 Speaker 2

Work on your bot and then when you're ready to test it, then you can test it on Dev and we can move it to production.

00:31:58 Speaker 2

Any questions?

00:32:10 Speaker 3

I have a question, yes.

00:32:13 Speaker 3

So can you talk about how this interacts with the other tools that we use and when you make a decision about when to put something in here versus other places?

00:32:24 Speaker 4

So this is it. Like I said before, a data warehouse. So it's basically a place for us to put straight up.

00:32:33 Speaker 2

Your data raw data that we've collected from.

00:32:37 Speaker 2

Facebook or Twitter or Reddit or, you know, Instagram or LinkedIn or what have you for later use in doing analysis.

00:32:49 Speaker 2

MISP in the Hive our analysis tools so MISP specifically is for helping us track patterns and find correlations between.

00:33:01 Speaker 2

Different incidents, right?

00:33:04 Speaker 2

So this person posted on Instagram 17 times and they also posted the same content on Twitter.

00:33:11 Speaker 2

So that lets us tie all those individual points together. Deccan is where we throw the raw data related to those incidents. So the artifacts specifically.

00:33:21 Speaker 2

At least that's the intent, right? And the reason why we're making that delineation is because a it, it's inefficient to do that sort of work in MISP. It's not designed for that.

00:33:32 Speaker 2

And B it lets us have the data that we can then go back later and do more in depth.

00:33:38 Speaker 2

Analysis on so.

00:33:39 Speaker 2

I could say grab everything related to 5G.

00:33:43 Speaker 2

Even if it's not related to just one incident, if I want to just go grab everything related to 5G or everything related to anti VAX or everything related to election security.

00:33:53 Speaker 2

The other reason why is because.

00:33:57 Speaker 2

We want to be able to share data with other teams, so we may have teams that are working on this information response in the UK or in Germany or you know.

00:34:12 Speaker 2

Brazil or Argentina or other country, you know other places.

00:34:17 Speaker 2

And we might.

00:34:17 Speaker 2

Not want to share our MISD or cortex or hive or cortex instances with them, but we might just want to be able to say here's a link. Go grab the data and so.

00:34:26 Speaker 2

That is another example where we would use Deccan.

00:34:30 Speaker 2

That makes sense.

00:34:32 Speaker 3

Sure. And then there's an automated way that this interacts with Miss.

00:34:36 Speaker 2

Not yet. That is a because of the way that the API works. That's something that we could set.

00:34:40 Speaker 2

Up, we just haven't done it yet.

00:34:42 Speaker 3

Got it.

00:34:48 Speaker 2

Other questions? Sure, go ahead.

00:34:48 Speaker 6

Grant, can I ask the question about the workflow? The notion is that I have my spider running, I grab my data.

00:34:58 Speaker 6

I decide that I should put that.

00:35:03 Speaker 6

And deccan?

00:35:05 Speaker 6

Or I could build a spider and have it using the API to dump what it's what it spiders directly through dcam.

00:35:15 Speaker 6

What is your thoughts about those options?

00:35:20 Speaker 5

Yeah, actually I.

00:35:21 Speaker 5

Wanted to add on to that. I was wondering about streaming data to Deccan.

00:35:25 Speaker 2

That is something we can do. It's been discussed, hasn't we haven't had a chance to actually implement it yet, but it is definitely something we could do. One of the workflows that we envisioned was.

00:35:38 Speaker 2

Basically, tacking onto the archive bot so we can currently in the CTIA league slack throw things in the Internet archive just by triggering the archive bot and then the URL. It would not take much to tack on the ability to add things to.

00:35:58 Speaker 2

He can by basically just saying, you know, either listing a data set name or A tag right. So we could say.

00:36:07 Speaker 2

Archive this in the Internet Archive and then take that resource and attach it to a Deccan data set and either we have a default one for COVID-19 or whatever the topic is or we say find the largest data set with this tag and attach it there right? Or find this specific data set and attach it there.

00:36:27 Speaker 2

Or if you can't find anything, just.

00:36:29 Speaker 2

Attach it to this catch.

00:36:30 Speaker 1

All data set stuff like.

00:36:32 Speaker 2

That so it's entirely feasible we just haven't.

00:36:34 Speaker 2

Set it up yet.

00:36:40 Speaker 8

OK, my question.

00:36:44 Speaker 8

Is this a good resting place for the data that's being compiled by the minions?

00:36:50 Speaker 2

I don't see why not.

00:36:53 Speaker 2

I mean, I wouldn't necessarily do it hourly, but you know like.

00:36:56 Speaker 2

A daily.

00:36:56 Speaker 8

No, no. But I can see doing it in a batch and automating it with an API or something, right?

00:37:01 Speaker 2

Absolutely daily.

00:37:03 Speaker 8

Right now, I don't think anyone's really using that data, yeah.

00:37:08 Speaker 8

And it's been collected on a regular basis for about three or four months now.

00:37:13 Speaker 5

Which date are you which date are?

00:37:15 Speaker 5

You talking about specifically?

00:37:17 Speaker 8

There's a series of programs I wrote called Minions. You can see them in the.

00:37:24 Speaker 5

Like like capturing not fire hose but capturing terms, yeah.

00:37:28 Speaker 8

Not. Not quite it I'm I'm missing a word in my vocabulary right now GitHub. You can read about it in the GitHub.

00:37:35 Speaker 8

For the for.

00:37:36 Speaker 8

The team, what it's doing is there's two different ones running right now. One scrapes Twitter looking for specific.

00:37:43 Speaker 8

Hashtags that people have asked me to look for.

00:37:46 Speaker 8

The second one uses a matrix of each state. It's agreed postal abbreviation and its capital with a prefix of things like free this state liberate this state, reopen this state, and so on. They're.

00:38:02 Speaker 8

They're two separate databases and what they do is they capture.

00:38:06 Speaker 8

Tweets they tracked down all the hashtags that are in them both inside of the username and in the text, and then it reads them for further research so that we could take a hashtag and take it down to the granular level of the numeric ID of the individual in Twitter who's posting it.

00:38:26 Speaker 1

Yeah. So that's.

00:38:27 Speaker 3

A ton of sense to do here if we could run it, you know, once a day or something like that and then be able to pull that information out and say.

00:38:34 Speaker 3

This is how that hashtag is trending or not, or whatever.

00:38:36 Speaker 8

It's already doing that into our slack channel.

00:38:40 Speaker 3

Right. I've been checking that out, but it it's hard to like make sense of that, right? So if we had it in one place in.

00:38:47 Speaker 3

Yeah, it would be.

00:38:48 Speaker 8

What's just just so like maybe I can fix it. What? What's what's the difficulty?

00:38:53 Speaker 3

Ohh no, just the changeover time piece, right?

00:38:57 Speaker 3

Let me be that.

00:38:59 Speaker 8

Ohh, change overtime. OK. Yeah, no, that, that, that it doesn't do that just gives you an update as per each run.

00:39:06 Speaker 5

Well, but you can do that once the data is in dcan and and and and maybe it's worth considering the workflow pipelines that we're creating in slack if maybe that's another command.

00:39:09 Speaker 8

Right, exactly.

00:39:20 Speaker 5

To add new terms so that you don't have to do it by hand or what I've also heard too is that if we're going to do stuff like that, maybe there's a workflow queue.

00:39:28 Speaker 5

And I suspect that the the D3 guys might be able to handle that where there's an approval queue so that we don't have people flooding.

00:39:36 Speaker 5

Lists or bots with extraneous or errors, so that someone can add it and then an approval admin can go and approve all of the additions to the scraper, right?

00:39:48 Speaker 5

So that you don't have to like take requests and do it. We can automate that. But then have it moderated through a A queue.

00:39:54 Speaker 2

In this case, you probably make the most sense.

00:39:57 Speaker 2

For [Name redacted] to.

00:39:58 Speaker 2

Do is we set up one data set that is just the the minion?

00:40:03 Speaker 2

Input and then we could set it up to basically run a Cron job where it takes that data, dumps the CSV into that data set as a new resource every day, and then we could do more visualizations where we.

00:40:15 Speaker 2

Could, say set up a.

00:40:15 Speaker 5

Like an ETL.

00:40:18 Speaker 2

We can track the data overtime we can see.

00:40:21 Speaker 5

Yeah, right.

00:40:23 Speaker 5

Yeah. So you can see those trends and also the longitudinal trends that you don't necessarily see.

00:40:30 Speaker 8

Yeah, see.

00:40:31 Speaker 5

Change the windows, but you know the the data is valuable. If you're if we're capturing it and and and to the point of being able to add more terms to it in a more automated or.

00:40:45 Speaker 5

Process oriented fashion. Not that we're not doing that today, but.

00:40:50 Speaker 5

Eliminating the need to have someone with, you know, a developer do it versus something the prover.

00:40:54 Speaker 8

Well, I I haven't had any. I haven't had any requests in like 3 weeks or anything.

00:40:59 Speaker 5

Well, I know that's what I'm saying.

00:41:01 Speaker 5

There might be.

00:41:02 Speaker 5

More because we're well beyond the open stuff now. I mean, we're we're back in open territory, but a lot has happened, right? There's a lot of terms.

00:41:08 Speaker 5

We could be adding to that that.

00:41:12 Speaker 5

Tool and I'm not sure since you're scraping right, you're not necessarily butting up against.

00:41:21 Speaker 8

What limitations?

00:41:23 Speaker 5

Like Twitter API limitations.

00:41:26 Speaker 8

Oh no, I get around that in the code. Actually, Twitter Twitter tells you in the headers that it returns how far away you are from reaching the limit. So the code actually interprets that, and when it reaches the limit, it pauses until the limits passed and then resumes.

00:41:29 Speaker 5

Yeah. Yeah, right.

00:41:43 Speaker 3

Right.

00:41:44 Speaker 5

And you could ostensibly, I mean, what we don't want.

00:41:46 Speaker 5

To take it off track on on.

00:41:48 Speaker 5

And wish listing this but.

00:41:50 Speaker 5

You get where I'm coming from.

00:41:51 Speaker 5

Right. I mean these are things maybe to think about as we think.

00:41:54 Speaker 8

Oh yeah, no, they they guys said that that's already been handled so.

00:41:54 Speaker 5

About the bigger picture, yeah.

00:41:57 Speaker 3

I think, yeah.

00:42:01 Speaker 8

If there's anything you want a wish list, you know certainly give me a DM and I'll see what I can do.

00:42:07 Speaker 5

It may be worth having a separate conversation about like.

00:42:11 Speaker 5

Or a tools. So I'm also trying to see if I can find some summer interns for us since school's ending.

00:42:18 Speaker 3

Nate, it feels like this is maybe a conversation for the managers.

00:42:21 Speaker 3

Yeah, like workflow and like how.

00:42:22 Speaker 8

Yeah. Yep, Yep.

00:42:23 Speaker 3

You want to manage that stuff.

00:42:25 Speaker 8

My my initial question was just whether or not the data I've got is a good input or absolutely.

00:42:29 Speaker 3

Yes, I think that's 100% true, yes.

00:42:31 Speaker 5

It was. It's been for the past three months.

00:42:34 Speaker 5

It has been already.

00:42:36 Speaker 2

It wouldn't take much to to.

00:42:37 Speaker 2

Set it up to put it into the data.

00:42:38 Speaker 2

Set so we.

00:42:39 Speaker 1

Can then run the.

00:42:41 Speaker 5

Analysis On it you can have an ETL transport capability where it can go out and.

00:42:45 Speaker 5

Grab data from other.

00:42:47 Speaker 2

Yes, I don't, yes. So it's designed to do aggregation from other open data sites. So you can do things like grab.

00:42:57 Speaker 2

The initial use case was grabbing data from HHS and being able to.

00:43:02 Speaker 2

Yeah. So as you can use.

00:43:03 Speaker 5

Yeah. So it does all that, ETL, all that stuff good.

00:43:06 Speaker 3

Yes. Can I toss out a question to everybody else who hasn't had a chance yet to go like [Name redacted][Name redacted][Name redacted]?

00:43:15 Speaker 3

I don't think it's [Name redacted]. Maybe [Name redacted]. Anybody else? [Name redacted], any questions about [Name redacted] before we? Because we're kind of sounds like we're transitioning it to happy hour ish pretty soon.

00:43:33 Speaker 7

But once going twice.

00:43:38 Speaker 3

Cool. Just wanted to about the space.

00:43:41 Speaker 7

Thank you, Sir.