Recording: 1 Feed Ingestion walkthrough step fn, lambda & python code 2.

the ETL feed bucket the error bucket and the process bucket so folder needs to be created and all the four bucket so that path error should not come right ones that things are done then we add the specific conversion entries like what kind of file that venture will be sending the respective Lambda function we need to create an entry like I have done for this vendors ke vendor ID paste something so what are the columns required and what kind of file it is what kind of separation it is using you know just an example type so we have this vendor account Id combinations I was talking about like this so we have to add another entry for that winter how that vendor which vendor file and how it is going to be converted ok all these needs to be done and then again the requirement because all friends ascending different file write the file types or the column types my default you know so for that we have to separate entity is there no grouping for the same like for csv file types or text file types of simple flow but yeah for the text file is because sometimes the file is simple like this hatched file so not ok so like this stage file so I was generic function you just specify the required columns which will definitely want from the vendor and we expect them in the file specified that based on the file because file may sometimes in txt also we have separated by space separation by semicolon separation by comma to the you know ok so that we need to specify the generic function will take care of it sometimes there are specific things required like this hatchet UK so in hatched UK that I think some extra things is required so I haven't add on function here in this we have different different functions like for hatched upi must be having one thing and then converted the file awesome transformation must be required to usko Karke and then I have converted the site needs to be take exactly exactly I think the XML if the things are not changing then it is fine but yeah for the other windows at least that totally depends on how the file is that's the main thing matlab how the file is how the columns are how well structured or not well structured the file so based on that we need to decide like how we gonna code for that if that nothing much to do it simple things are there just call them accordingly and you know you can process it if things are complex in the file within the file then we need to handle those things before done at the lamb conversion level is done now we move to the ATL part and create a job there according to the vendor specific columns and mapping in everything based on that that I'll cover late tomorrow and I'll tell how I have done sometimes I have kept the same job for multiple Windows the grouping you were talking about sometimes I had to create a separate job because of again the file and how the columns all these are done then we need to update the system conflict table which job name we have created for that ID ok how many concurrent jobs because like for in gram I have multiple concurrent job is too much so the same code is there the same file structure is there so I can run multiple concurrent job for in gram I think I am running 20 Parallel Lines so that way current job thing is there and third is the path which folder location the file needs to be fetched from and looked into this path high these entries are being added in the system config table the system conflict table that is from the backend from the

database and that's not from the you why you need a developer to code for the conversion the same developer will update the entities entries in the system conflict table from the banking this is not provided in the US because it is not for the regular user this is for the development of you right and from there so you will create a lets you created a job in the interior so you will have a job name you will create folder path accordingly in the SC buckets so you have the path and the concurrency you decide based on how the frequently the vendor is sending the file or how big or how much files you know it is sending like in gram like generally to conference of the sufficient for all the windows were sometimes it's a big file to it get to sleep is splitting into 25 26 for the 26 and 1 by 1 to process it will take more time parallel processing so that it is indicate ingested in the Staging table and then the further the process can happen so that is getting injection in the database I will tell in that tomorrow session when I will take up the ethical part these are there these conference jobs are like 25 I have given so it's mostly remains below 25 so that it should not exceed the because again if it exceeds it will go to hell I have given limit kind of things 25 super Nahin Jana chahie so the system the Java code knows how many parallel processes to be you know started for that particular thing how many step function 2423 parallel Threads you know for income parallel and then the same conversion blue vitriol and that same flow will happen for all of them simultaneously that so that it is required so this is mainly used in the Java code but we need to specify and set accordingly in the ETV and feed system config we have another catch as well because of this certain you know the same venture and multi accounting so sometimes we had to have this ideas like difference so we have different names if you notice right so if it is written as ETL job ETL concurrent job and feedback so that means the key value provided here is the vendor ID ok if it is ETL jobs by account it will contribute job by account and feed passed by account that meet the value specified here is the account Id and there was a case where we have to refer the field ID because at the feet level a particular FTP is changing the account is same in the vendor is same you know so for that case we have this by feed so it will job by feed it will conference it passed by feet and the value specified here is the feed master ID of that particular account what is this conflicting right ok initially when we started the call explain the three Masters right vendor master account master and feed master correct jobs and feedback these three entries to be provided for every ETL jobs ETL concurrent jobs in feet the information has to be passed for every new vendor so if it is written like this just like this ETL jobs concurrent jobs and path so this three this three is written this three is the vendor ID this makes the collection of these three number this number is associ if it is written like this it is job by account concurrent job account bypass so by account means the number specified 18 18 is the account master ID even backtracking right also we need to know this number 18 is coming from which master if it is written as 5 ft matlab feed master if it is nothing is written that means vendor master ok so when backtracking this will help and when creating a new one so whatever the scenario is you know we have to create accordingly so we are we need to take by account because sometimes happens for the same vendor the file type or the type of columns or something is different for two accounts of the same window in that case have to specify by account because vendor ID is common for example if a vendor let's say this

GPS GPS having multiple account ok so GPS account vendor ID is 4 for example suppose we have GPS which is the vendor ID four ok and GPS we have GPS UK account Id let's say what to house may be right so for example this is scenario right now what will happen is if you create the jobs by feed so for all these three values the vendor ID is 4 correct now what the file type and the type of columns as sending its different for each vendor in this scenario instead of taking 4 as the ID here in the system config we map them by account for account Id to which job to refer which part to refer how many concurrent jobs it can have for account Id 12 you know same and similarly for Accountant in 24 in that scenario this by accounting comes into picture so there is a lot of when a bigger vendor so GPS is a vendor ID for now what will happen if I do is K if I create this job by this way if I create this in this manner and I put for hair ok so what will happen is if the GPS UK file will come it will send to the same path and it will refer the same material job if a Crown house file will come it will do the same thing and if for one page file will come it will do the same thing because the vendor ID is same for all three accounts because these three are associated with GPS correct right now what issue is a GPS UK sending a txt file Crown house is sending some XML file and cognitive sending XML file but some different columns or different values so this four will not justify right the conversion will fail for either one of them are the ATL will fail because the columns will different right so in this scenario we have to go to this account by accounting so for GPS UK right right so for GPS UK will do this we take the account Id two will put to here and will pass a separate job for GPS UK separate folder path folder files for Crown house will do the same thing will create this thing by account and put 12 and we specify a separate job so that they won't overlap with each other if the file is same if these three vendors are sending the same file then doing it by this but if they are not doing then we have to go at the account level and do it by accounting then right and so clear with this one because I have to tell another one go to go so this by feed so now so many condition in which will happen for the same thing right now I understand working on this for 2 years now I understand out right right comes into picture like I think I remember a case for cotton we had to do I think it was the one where it all started so for cordova sending simultaneous file in his FTP so Simon sister we have separate vendor separate account the file is for the same window same account but coming in the court of FTP along with the semester FTP as well as sending sending the file for semester result and the final team wanted to be interested for semester obviously so far in that case we had to introduce the by feed method and be referred the feed master and we pull the field master ID ok and then we referred to the left which accounts it is mapping but the ID referred is that master but again the accountant vendor should remain the same area for this by feeding ok clear sofa so we have covered this portion this is a step functions and this is the conversion and the Lambda and tomorrow will be focusing on the second half of it that is the blowjob and the data entry so which tables are being used for the injection part how it is happening how the procedure is triggering and all those Part I will cover tomorrow any questions so far you you answer the questions that we ask but still whenever we will come for tomorrow session you must be couple more yourself go through this process in the initial of the call you explain something related to the error Lalita and the team is facing

explain I'll explain that it will solve the case anyways so I think what happened is UK right so till 19 July everything was fine from 29 July onwards the coronavirus UK started going to error and we received the mail this is going to add please check so when I was evaluating the first thing I checked is the masters after like because when I check the logs I couldn't find anything they referred to Masters and so Crown house me I think crowners UK some new vendor is created Crown of UK I think earlier it was under the GPS to the account Id in the vendor ID combination we were using words of GPS and criminals UK now the Crown of UK is associated with Crown of UK a new vendor so I went to vendor master round so this new vendor is created right the ID card change to 542 so now my system is not aware of 542 and 18 number combination show the Statue of the process you followed through that this vendor master and account master ideas are very crucial because we are using with them throughout the code to identify which vendor file we are receiving and how to process it was this is the only way we had to distinguish between the vendors when the file is getting to the system so if you want to change we cannot do that we can do it but some other changes along changes are required before the successful in a processing of the file again since this change was made I think on 25th July and from 29 July onwards it started failing because this combination is new to system 18542 in a system don't know it will say the you know vendor for this file type is not there and it is just you know giving an error so this is how crucial the account Id and vendor ID and this is specifically when we are talking about the incoming Windows system hardly you know matters whatever it will associate and give it to him it will process it because manual field will be our separate flow manual file whatever will come it will come in a CSE format we have all the columns there is a standard essay of the vendor account Id it will process the manual feed data but in the incoming feed the file types are different so we had to Cate by ID and identify which file type is coming what in the txt also if it is sending so all the vendors are not sending the same txt file in the same format XML and that even the txt files are not same for all Windows either the number of columns might vary with the column names might vary or even the partition you know the space or semicolon or separation also might very crucial to have this account exchange if I talk about the problem with this it was a new window was created but the system did not have at the information about that like what id does it have correct and due to that the file the field that I was getting through this window was also different because you just change this association because your business might require right now I don't know whatever it may be but yeah whatever name you are refreshing in the invoice you can put here very easily so that your team can identify you can rename the accounts according to your convenience and but just don't interchange them very frequently because the code has to be updated along with your changing the vendor invoice name or the reference name of the account or the vendor itself ok system is fine with it but when you create a separate vendor along you know with different ID and everything then things then all this information either to be remarked in the system throughout the code level where it was required wherever it was required Hai vahan per changes Rahenge and in fact this will also affect the earlier data for example let's say GPS so the GPS was the full name WhatsApp in which we got initially has an account identity 11 and it is associated with

groundnut account Id 18 so 11 and 18 so when you check the data in the advanced search right and you select a vendor account basically the vendor account account Id you didn't change it is still 18 you will get the records associated account Id 18 but what will happen if the previous generation was with grand them account Bend It 11 so you will get the name vendor name as grantham book service then once you changed then you will get the new name for vendor that will also happened data was there has an association with vendor account Id and the SBI sorry that's why that is also the reason mainly that inform the team about this K whenever you are creating a new vendors tricky it requires a lot of changes from the backend along with the front end coming I think I am good from myself will be going through the same thing process from iron and tomorrow when will be having your session first we might whatever doubts we are having and then we can proceed with the jobs in India no but now I am good so thank you so much for now thank you thank you the day tomorrow

Recording: 2 Feed Ingestion walkthrough step fn, lambda & python code.txt

that is because of that reason only change vendor has changed so suppose it is before it was built from white something and right so I clarify is that will discuss over that separately from so initially what happened is the time when we started we have a certain mapping to certain vendors so we categorise the big vendors and we take their names are the return percentage and the country code and all the information in the priority importantly to the main vendor a vendor is kind of a bigger entity for example consider the accounts related to value right so this is the accounts which are related to widely and for the vendor invoice name we have separate calling here right so vendor invoice name if it is changing and separately also the vendor source vendor is changing then only in that case will change the vendor mapping if just the invoice name is changing we can update the invoice name that will not affect the code Association is required at the time of you know the injection that is very critical because these ideas are being used in the code for example right now I think the for the program is mapping got changed and because of the that the other places like the system and the code level at the DB injection level please mapping or not changed from the back end so that is what and account mapping is critical specially for the incoming vendors for the manual it is fine you can do changes and you can upload to it but for the incoming vendors these ideas are very critical test if the invoice name is changing then it is fine yeah Mayank is waiting can you please yeah so this so whenever we decide to change the vendor account mapping in terms of Masters we need to take care like the IDS to be updated in different multiple places along with these changes throughout the code and

the data and everything you know that I didn't need to update you there as well because we have a couple of tables you know which injection is taken care of this one system configure system configuration you can see the numbers write these are the ID of the vendors and these are the different Jobs or parts related to that id if you change that I did the system will not identify the correct path of the correct job to be taken care and then it goes to error this is one part of the places can you help me out with how how do we add a new vendor if it's a new vendors join the system adding a new vendor is not task for the incoming feed you are asking you get a new window they want to get into the system so what's the entry point So there are couple of steps to be taken its new new entry in the vendor master then after that we need to create from the vendor master you can update from the UI then you need to create an account from the UI and you can map associate with the vendor master in the account itself when these two sets are done then we need the specific codes write the glucose and the conversion code respectively which file in the vendor is going to send XML file or a text file or that file or csv file what kind of file vendor will be sending based on that we need to you know make some changes in the conversion code and along with that we need to make some changes in the glucose based on that file type and the columns which the vendor is sending the mapping of the columns what mapping is going to send What data is going to send those entries we need to create those entry and then we need to update the system conflict table with the respective Ids and so so please listed in your I mean like a point to be told like just walked through the same so that we get the clear idea of how the process works right so I will be telling you first the entities where were all things are and then I can get all the time thank you for what is again that is not very critical so this is the one of the services which are you know having all the graph you can see the flow to these are the step functions in step functions step functions we have state machines so in this we have this only one shake machine that is feeding machine step function and open any one of them so this is the graphical view of the flow all the green parts are successfully executed and you know the processed to in this step function we have associations like we can associate different different other services like in the conversions if you see the first one ok so after start you see the conversion this conversion is connected to the Lambda functions which contains all the code for the conversions ok once the conversion is completed if it returns success response it takes this thing and then if check if the crawler is required because initially when we designed in the crawler was required to be had this thing we had this crawler well before the clothing now after the things changes the upgraded version came so that floral was no more required to be added to check if it is optional thing because we were in transit still there but now the value is sent null so no problem is required then it

goes to the Glue in the Global transformation of the data happens and post that checks if the trigger migration is immediately required or else it needs to wait so Trigger migration is related to the database entry part ok so today will be covering the conversion part and tomorrow will be converting the blue part along with the induction in the DP ok so what happens is when a new file comes to an FTP location the FTP services which are made in Java monitor the FTP servers and the file if it is not the file name is not there in the database it picks that file it creates an entry in the feedback table in the database so it will create a run ID it will create an entry for that file which file path through associated this is alone from the system table again what is the vendor ID what is the field master ID because for ingestion we have feed master table as well I will come to that when explain the new item name what's the original file because what happens is sometimes the file size above 250 MB we are splitting the file into multiple jumps so in that case the parent file name we expect and we have the individuals chunk name here as a feed file 1234 like this if you can see on my screen 25 24 23 this all the big parent file name is this song so this was the single file which came on the Ingram FTP and I think it's somewhere around 1.3 GB zipped and 3.4 GB once it so it when to 26 Jan it will be created the new run ID all the parts the vendor ID the master ID the identifiers in all account Id so this mappings are very critical if any of this changes then all the you know system gets distracted their current status if it went to retrieve then what is the retrieve account if it fails what is the remark migration and diplo migration statuses are related to the database entries how many total records got process what was the upload type it was injection with the listener injection manual means it's a manual uploaded by listening and free type if it is a p and BBC error records if any how many of them went to wear and all these things are in the feedback table so first after the entry is created in the feedback one JSON is shared with the step functions as an input go to share we share the file name the input path I'll show where it will use input bucket so I'll tell this again so input bucket we have vendor ID account Id source country crawler if required then the name of the crawler if not it's null job name output bucket process bucket Run ID migration if needed immediately vendor priority what's the the vendor what's the feet type and who created all these fields come as an input to the step function and based on this input the further process takes place ok so this input is passed through all the steps the step function start how do we decide the vendor priority discuss with the team while deciding the priorities I think one is for the manual upload when it is you know not to be overridden that is the top priority to I think it's for the publishers 345 it's like for the Ingram focus in garden order I am not pretty sure right now I don't remember exactly somewhere like that 345 is the income then the other distributors and then the Indian population in India distributors this is how the

priority level is there right now and the code is like that where are these priorities are affecting only the quotation for this priorities are effective initially in the DP so that which data can be operated in the bibliography so which vendor can modify the existing data that is been decided by the priority so one can override any of the data to can avoid anything except one that way ok ok to that any vendor with the letter should not update the priority more priority vendor information that's right there it is used so suppose if we change any vendor priority manually in front end so it will affect others things also or we need changes in the back and also for priority we don't need change in the back end I think if you change the priority to some other value so let's suppose for make it 4 right so what will happen all the records updated by we still have the priorities that is the issue and if you make it 4 so new invention will come with the priority 4 not an issue but it won't might not be update the priority 3 existing values in the Bibliography table so that is the issue break the code but it won't update the values in the Bibliography so we have to run a query specifically for that vendor by adding a well whatever the new priority we have set to update the existing fields for the same vendor you know so that it can update the priority this will happen only if the priority is made from 3 to 4 if it is made from 3 to 2 then it is fine 2 can update 3 so before moving to the conversion functions let's see this S3 buckets because we are using couple of buckets here for the ingestion part explain each one of them so the first packet that is the data insertion ETL this bucket contains the code part on the codes which are stored on Amazon and stored here in this bucket data insertion yeah ok so we have a respective folders the ideal job codes are in the ATL folder under libraries we are using that is in the Lambda libraries folder add on libraries is the zip kept here itself and on its code mappers are kept in the this folder so we have separate folder for the separate jobs and all the code level part are stored in the data insertion ETL OK then we have this RAW feed files fraud this we have a separate folder vendor wise vendor account wise ok which vendor account you know we are dealing with we have a separate folder in this folder further we have a p and a folder and in the pnf folder all the raw files are uploaded which will get on the FTP like the txt file is getting will upload it to the raw bucket is named as raw so and the buckets name will change the path is constant on all the buckets we have the same path in gram DNA and further file name whatever we are having so in the raw bucket the file will be uploaded from the raw bucket once the conversion happens the file is moved to the ETL feed file from this is the active bucket we call it because in this only file is kept when it is being processed not before not after when it is in the process the file is kept their and once the process is done irrespective of the failure or success the file will be moved from here ok again like you mention the raw files will get uploaded over there and then the processing will

happen and till the time of processing the file will be over here in this file conversion after conversion so once the file is uploaded in the raw format and the conversion happens for that file and the file is stored over here so here also you will be having all like vendor specific folder names files or how it is Right exactly the same same structure same structure we have because in and then DNA and till the time the file is processed OK it is failure or success the file moves from here on its gone to the processing part right so if the file is process successfully we have a process feeds bucket solve this file is completed successfully no error nothing the file will go to this bucket same folder structure because in PNA and the converted file process file will be there in the csv format was ultimately converting all the file to csv how this number is being generated after the name that is all the information this time stamp complete file name you can find them and if for any case it goes to error after conversion not before conversion it goes to convert before conversion it gets to error then the file remains in the raw bucket won't be passed to the process anyway right after conversation the file into arrested and say again if you go to conversion so it will take file again or it goes to some error position or how you get to know that this file needs to be processed again or not this needs to be skipped ok so what happens is anything which goes directly to error in the UI if we see this data field incoming free type so we have this status pending is not yet process speed started if it is error that mean it has failed in the conversion layer itself something issue with the file should be directly check the file what is wrong with the file it remains in the rocket was it not yet converted only if it is into retrieve that means the file is converted successfully but something is wrong with the ETL that glue live then we check that part if it is in progress that means the process is going on not yet updated in the DB are in the probably in the Glue state itself somewhere in the process once everything is completed we put it to completed the status we decide which level it might happen and we assume if it is filled with the before the conversion itself that means something issue with the file there is no point read trying it will take the file right so if it went to error we have a error feed file bucket same folder structure DNA and that error file we can find it here in the csv format so then we can check download the file we can you know all debugging part we can do what was the file what we got in the file what failed in the ethical part you know which check the logs and everything and then we can you know pics or conclude accordingly these are the four buckets and all the names are passed you can see here now again so this is the path which is common file name which remain common extension might change so we don't consider the extension we consider the path name before the extension the file and the bucket names input bucket that is the wrong one ok is the ATL with the processing bucket process if it successful then the process speed and error bucket uske liye energy

files which is backend mein bhi storage no need to pass the error bucket name with directly that is managed in the back and all these parts are created yeah also find the consoles for the those error files like how do you check the consoles for the error while in only limited the locks for one day to fix the recent issue we can directly go to cloudwatch we get the whole you know login everything some basic logs or something we find here itself at the level so if anything is fail we can click on the particular layer suppose convention we need to check the click on the conversion layer and in the output we can see some logs at least so we get a reference point sometimes there and if not we read in the job and see where it is feeling so that we is generate a new law suppose some job Field 3 days ago I won't be having blogs and probably filled in the Blue Label so don't no loss will be shown here no much information is not there right so I proudly read on it so that again it goes to error then I can check get the new locks and I can you do that because so it will be covered tomorrow so I'll tell you like this whole step function so what happens at this new execution right can keep the season click on the new execution after opening a particular job I will get the JSON as it is I can Garden name specific test or something so that I can know identify this thing and I start execution on the execution is start it will cover all the layers step by step to conversion job and everything will happen again for the same file from the raw bucket because the raw bucket means we still keep the file anyways right what will convert again and all the process will happen so you can you see at which level it is failing and what level and the thing is going wrong older than one day if it is within 24 hours we can directly go to watch cloudwatch logs and we can check one more thing I don't know it's related to over this or not face like the feet stop getting ingested in between sometimes it's not it's kind of related it's not related because what happens Sim induction is happening with multiple vendors and the sum of sending large files in quite more directing 67 of them in ascending large size so what happens at the peak cards it creates kind of a bottle milk so restarting the server clear that bottleneck and the injection start again you know you can understand pic art this is happening when a large files are coming simultaneously for some vendors it is creating kind of bottleneck and it my affecting some vendors injection Listener Service is there if you restart that particular service but it will still suspend the all the things that could have been running restart will Listener Service is responsible for taking the file from FTP and putting it to the step function so all the job which are already in the step function in the running state that won't be affected that will run as it is so it won't happen this part which are already in the process it will take up the new ones and it would hardly takes I think one or two minutes to get the service up and running again so it is like very negative so this way we can proceed now we will talk about this conversion layer so conversion layer is

associated with the Lambda functions rate opening location [Unintelligible audio] consider the same case like for Instagram this file came in the raw bucket came to the conversion layer another function this conversion layer is calling is this one parent conversion function it's a Lambda function return in Python refreshment code is not upload so this is the main parent conversion function which is associated directly the state function here conversion layer ok the file goes to this function and this function file come to this function and we have this mapping for the extension file extension it comes to the function it check it breaks down the file it gets the file extension with extension it is there and it checks suppose for Instagram the extension is so it will pass this job from the conference conversion function this function itself will call the XML to csv function the specific conversion function 5 and that to see the path specified because this is configuration layer taking time so I have created dictionary in this this is the dynamic path has given that is itself in the parent conversion function so these are the part to send other function which I have created in Lambda the HTML to csv that to csv text to csv and I have associated this part with the extension so which ever extension comes in gram it is coming that it will automatically call this path which is this that to seriously and then the that to CC function will convert and process the Ingram file return response to the parent conversion function move the file to the ATL processing folder and gives a success response to the step functions right so in this this is the diet to CSE function we have a Lambda handler function is the main function which is called this is the main function which is called when we calling Lambda and in the event we can pass JSON inputs to the functions and everything and here I can get checking so this is how we are checking the vendor ID and account Id is so if we change the mapping there and you don't change it here the system Windows start feeling goes into error give you some different different this is the case this is a very important in the code level starting from the conversion tell me the ATL level and further in the DB as well ok so if the vendor ID is 3 that is outstanding that file to in the same file it is being processed if the vendor ID is in 8 or 451 that is income data and IPS data and then I created separate functions for 1 gram file as a separate function you know customise for 8 gram you know columns and everything with column mapping and all because in that file it is not sending even the column names it is distance based file creating the column in the distance at till which length a particular value belongs to at which column this was all in the Ingram document shared and then further you know whatever conversion required we are doing it like for hyperconnect USB had this vendor id3 I was separate function the data is passed through this function and then the conversion happened for that particular file a particular account the content is written function back to the parent function I was telling you that your requirements mentioning like we

don't want few of the indexes of a particular vendor to be like overridden or even changed or something like that so where that part is being held if you want to make any specific changes to a particular vendor with the incoming like an example can you give me vendor you are just mention that coming and we don't want the cost part of that particular feeds values to be updated when the new fields are coming all other index of the column should be updated but that just that particular column which is having the cost of that particular has been need not be updated with the new incoming field so how we can customise that for a particular window ok so that can be certain level Tak for other end of the indexes we can handle that in the ETL part we just don't map that column so that value will get as null in the DB and in right so that way we can prevent that particular value updation for prices it is that is tricky because prices is notice you know someday you never know we are not updating so there we need to pass zero something to be done other than that any other column a value we don't want we can you know remove the mapping in the at the ATL level so that won't column would be added it will go as null and null won't be updated in the system that will remain without converting whatever the vendor is sending in the exact format so that we can get the CST correct deciding what to take what not to take so I was telling in the parent conversion function we have this configuration in the environment variables have specified the path to the different different Lambda functions and calling these keys passes in the code so that I don't have to hardcore the function hall together switch on the extension it will call the specific function and in the specific function we have a specific code accordingly to handle that particular file and that is been decided just by the account ID or vendor ID or a combination of both sometimes because one vendor have multiple accounts so we sometimes require specific combination of vendor account Id is done to identify which particular vendor file has come and to convert it into a certain manner ok similarly for XML files we have XML to csv conversion with a similar structure and everything for text file we have text to csv and if it is csv file directly then there is no need of converting it will directly pass that file into the further process generic function I created this is the it is called Lambda handler so all the process happens here extension is checked if it is a csv if it is a manual field we have a separate in a kind of a flow other than that some experiences are not sending the proper extension the sending just the file name not the dot extension for that we have entered based on the account Id knowing the file type like for example paper bag shop ascending order CSC Mein send Karen but there is some conversion required for passing it to the function extensively you know extensively specified the ideas and I pass the which function to be called that way and then based on the vendor ID combination we have like op is not sending the extension like which extension to be called and the

generic flow is directly if it is an experiment go to text to csv if it is a that go to talk to CSE and finally consolidate all the responses and gives the success response to the step function again at this step and all each step we are having this notification kind of thing which is which goes to system in the status gets changed so this conversion completed will check if it is success of failure if it is a failure it will go to Royal Enfield and it will end the file at the floor itself raise exception and it will end this one has same if it is completed successfully it will check for the next step crawler required if the crawler is Null that means it's not required anymore then it will proceed to the glow step in that case if the file has gone into array so that would be moved to the error 85 all the row feet file so if it is filled at this level itself conversion right so in that case it will remain in the raw file because conversion hua hi nahin right the file is not converted to can't be moved to the processing bucket it will remain in the RAW and it will go to error if it is converted successfully then it will be passed to the processing ATL bucket and then if it fails in the it will be moved to the error record market the conversion has happened but the internet so in that case the file will move to the error bucket and we can BackTrack in your everything we can do we can find the file which is the converted file how it got converted maybe you know sometime garbage value is comes system is there so you can anything can happen slight change from the vendor site happens to the commerce anything you know sometimes if it is csv file or it's a text file and you are splitting it will comma and you know position change itself then you know it is a issue or problem vendor is added a new column so the column number of columns you know original something has happened so those issues are there because of the proper you know something or not getting proper values in that case we can back how the file got converted in the csv format and brought into success it goes to the process bucket this is a successful and similarly when we download it from the hair right so this download function what will do is till the process is in progress it will download the raw file completed it will download we have to have the files in the process feed bucket and the results in the robot boy and the error bucket for our own you know analysis and everything for debugging purposes and this ETL feed files this is just for the processing part what is the file is process whatever the result may be the file will be moved so this is always remain empty you don't hold the file here after processing ok ok so the couple of Masters for Speed Master so this is also very important master in terms of injection this is the sftp of the FTP information we are storing right this information goes into the field master with FTP host what's the username what's the source country again what's the time if it is FTP or an sftp server what account it belongs to so when we create a new account we need to add an entry in this master as well so we create the vendor in the vendor recreate the account in account master then we add this entry of the FTP

information in the field master these three information to be filled from the yuva itself not an issue ok so associate the FTP from with the particular account if but the password and everything and what's the file pattern what kind of file because sometimes when does tendency of sending multiple types of file in the same FTP location a problem we don't to put pull everything so we have specific specific you know this regular expressions we have what particular file or vendor what kind of file we need to put pull from that FTP location like in some cases think the vendor is sending US and UK both files in the same effect is also we need that file name or something along with the extension or sometimes just extension is sufficient need to find this file pattern is important and anything else is done just one more thing for the XML file types from the DB level we need to add extra information I'll show you what is the field master this one so any better which is sending this XML file we are taking right so what happens is in XML file we have this closing tag and opening time in the external structure sometime it is a small tag or short code one sometime it's a long time so we just need to specify the start tag and the entire what kind of tag this vendor will be sending this is from the DB not on the UI part so once the update the FTP information there we did start tag and tag and any other additional tag is there like the major get to know this tags to check the values so this information is need to be filled just for the this XML Windows only required information can we know manage from the just excuse me for a minute please sorry [Unintelligible audio]

Recording: 3 Stepfn_ETL_GLUE_Python_code_KT-1.txt

ok show me forward back to know Lahore last season one thing again I would like to go through that mapping process like you mentioned right account Id and delete ID needs to be mapped to know if you could do that particular part again [Unintelligible audio] show that mapping process is like that suppose we have because in a winter master with the priority number 6 is associated with three different accounts one is the Vidit 9 secondary us with 1853 and primary us ATPL account with the society of ATPL not PWD 193 it ok if we change the vendor invoice name that name changes for you guys it's to change the name ok you can change the gender if you like lekin rename the hair as well in a result is harmless just the association and understanding with the respective things that matters but if you create new one the ID is night change this is where the problem begins let me show you how to send txt files text to csv conversion rate now Bugatti primary us account that is the ID account list of columns listening and secondary us another compare ingesting that is 53 comes under 19696 and I start processing because

in that might give me error because the columns are different about the files so I created a mapping of a vendor account then you need to go into you know this section and check the required columns accordingly and then process the file and if the account Id is 53 along with the vendor ID is 6 then you need to check these columns with the and then process the file because the file things are also different like it is in the books in primary I just casually called that Pandas function in red the CSE symbol writing it was separated by tab instead of Commerce separated instead of the regular thing because of some issue I need to pass this engine Python I think it was feeling so this parameter is what separate things and specific mappings accordingly know what you can do is if you either change your vendor you create a new Window ID of 550 and the account Id is still 53 my system doesn't know this combination what will do is it will raise an exception as vendor with this file type not form need to write that particular combination of a hair long form need to check there also like we have used this account Id combinations these two places that needs to be updated along with any changes made in the UAE this is a crucial and in the system conflict table as well if required based on like how it is mapped as I should yesterday so if I see if for booking it is with respect to vendor account this you know normal thing and pass with the vendor ID can you change the vendor ID there then you need to update this as well or you need to create a separate entity to accounts are linked with the separate vendor ideas by account then there's no is that is fine clarity OK Google call highlight real job conference of the three parameters are map with the wind direction ok 6 create a separate vendor for the focus in secondary the vendor Id 6 won't be applicable on because in secondary thing right so system will not know which job need to repeat go that mapping is not then the system config as well so when the distance that process for step function it will not know which job bus 550 ID nothing will exist and for same for the 53 account Id so in that case what you need to do is you need to update this as well and make it by accounting can see this line number 13 by Account right and then you need to separated by account ID of the Google in primary hair for boys in primary job and for the secondary by account and then pass the account ID difference primary and secondary which we are talking about account level because they are two separate accounts linked with Bugatti it's more of a business point of you think we want to map with anyways need to do it it's not like fixed that if it is secondary you have to map it with Singh and if it is primary you need to make it like that in system configure three things are there like this is the default one ATL jobs nothing after this ending ok this means that the ideas which are mentioned here are the vendors ok 13 14 15 this is mentioned as by account when it's mentioned by account that means the ID which is passed here 4 this is an account ID account master and 4 is the Cop vendor ok that we can identify the compound if it is mentioned

by feed then it is the feed master ID this is the rate of this is a very specific use case thing but mainly used with the vendor ID or account Id so you need to look by the suffix yes what is suffix it's my account then make the ID mentioned is account Id if it is suffix is not there then it's a vendor ID if vendor is changing and things are changing so you need to make changes in the system config accordingly and in the conversion layer I need to check the three places you need to update accordingly if the mapping is there you know that way if you change the vendor account mapping anything else with respect to the floor we have discussed yesterday or any particular service or any configuration and any doubts we have regarding that is false anyone from bicycle also second half which is the ethical part please open the Indian service so this is AWS blue ETL here we have the transformation jobs mentioned and the further processing of the data and it's done here the mean two things one is the eight years of itself and second is data connections which we are using Ok so before creating a job what we do is we create a data connection with the RDS instance updated your database instances in the RTS so we have to establish a connection with that instance and configure this once this is configured now the Glue can connect with the connection and use the database if this is not set up properly blue will not be able to interact with the database this is this connection is important you can create a new connection you know or edit it and editing you can test this an option to test that connection as well so when you configure it completely with a test the connection so that you are sure to get is able to connect the blue is able to access the RDS that's very important once the connection is created with directly can jump to the ETL jobs because crawlers are being used but not in the process but for a different thing so what we have to do is let me come with the low voice so what happen is after the conversion happen the 18% start after the 18 processor we have three tables mainly which part is dealing with one is the feet Staging table who is the Bibliography table and one is the price in availability table ok so what happened is after the processing of the data the 80 year will insert all the data in the field Staging table now Staging table we have to store procedures which are triggered by the Java services ok and they respectively insert the data for the PNA data the price and availability which has a constraint of ispn and account combination is shared by multiple vendors right so in PNA we have this constraint can be stored the data in the combination of ISBN and an accountant so multiple time with different account Id ok those kind of information the price the currency you know the way the quantity discounts which is no different for different vendor or vendor account those things are stored in the p and a table and bibliographic information of the book in the Bibliography table like the title the author you know the population and this is just specific to one has been will be only available once in the table just giving you a novel

sold better understand like how the data is getting interested in all ok and complementary one we have another table with the error records so what will happen is at the time of ATL process is certain records you know which I found you know having some missing values are missing fields or important information or having some garbage value of any kind of issue I'll put these records in the error records table so that I know I can BackTrack and check like what was wrong with the particular records in this account is also available in the feedback table so you can see in the UI as well incoming feed the process account information the total records and error records is there any error records that will appear here the count and you can refer this error records table and see what went wrong for those particular records based on the run ID this portable are being used now the use of crawlers ok so after establishing the database is a data catalogue thing in the blue itself just created a panel DB in this way I have created two tables because these tables are at the time of injection these are used by the internal functions of the blue so data catalogue is table has to be there so what is the Staging table where I am in the proper records in one is the error record table where I am inserting the error code tables in the data catalogue databases I use Scrolls and cross these tables so that the columns and the data types of these tables are available in the system crawl the table in the database and then update this database in the catalogue catalogue is the column names and the data type of those columns so now after this step the Glue nose what are columns available for that particular table and what is the data type of that column while inserting its easier for you know understand and process the data Axis are not being used Akhilesh open multiple jumping and complex things data catalogue nothing in which they have their own databases and schemas and everything which they use at the time of injection ok now I am not writing any insert query when I am writing ETL jobs I am using the inbuilt functions of the blue itself ok so to insert the data the Glue have internal functions that needs you know access to this data catalogue data bases only in this databases are connected to the RDS itself indirectly I'll tell you how so we created connection correct establish the connection between RDS in AWS glue then we use that connection and run the crawler so crawler will nothing but crawl that particular table which are specified so in my case as specified two tables 1 is the field Staging table where I am inserting the proper records and one is the feed error inserting the error records OK so the two different colours will crawl two different tables check for what kind of column names I have there and what is the data type of each column after grabbing that information from the RDS it will update the database catalogue databases in the Glue so not glue will have this information ok these are the column names available and this is the data type of that column the internal functions of the Glue it will update ok this is the column name I was I wanted to

update and this is the column I am getting ok this which value goes to which column that's how the Global design so only in this scenario we are using crawlers that could not using products anywhere value product is not difference so what was happening earlier was I was scrolling the files as well the csv file which we are converting we were crawling those files as well to get the columns and data types of the column before investing it ok so this crawler in the diagram is for the files now when the Glue version got updated it can be directly it could CSC it could directly read the csv files and get the data types directly so the need of corolla was no more for me does it was an add on it was delaying the process anyways so we put it as because it was ongoing change in the production of directly remove with some things might fail in things were there so we kept this in the floor and we pass the value ok if it is selectively required for any particular job if we feel can we need to cross a certain things will pass the crawler name else it will be null and we put a condition that conditional directly proceed to the closer so we kept this option again open but you are not using it so this problem was for the files the problem which I showed you right now is for the database tables doesn't need to be Run every time just when you make any changes in the database suppose you added a new column ok probably you remove existing one or maybe you change the data type of any column any kind of the data type of the column in those two table feeding feeding table and in the error record in that case only you need to run after making those changes so the database tables in the blue it also gets updated accordingly that's update both glue and RBSE only only when you make any changes to the existing table columns of something 2022 August 16th September 20 to 2023 so whenever I made the changes are probably as per the requirement I need to add a column right whenever I added a new column I run this scroll down so the database tables in the data catalogue of blue you can call it new TV in short blue DB will get updated accordingly then it will be in sync with the RDX you need to run it is fine every time it's not don't need to run one all these things are done on the jobs come into picture so we're dealing with the book was an example again so Ok so are you can see the first nine lines these are the blue library in the which the blue provides the price for AWS blue inbuilt libraries and the functions which we are using in calling from the next line I have mentioned it as well reusable code libraries so the code which was you know again repeat the function which are required in almost alternate or every other job has created a separate files in libraries for it and I'll show you how I did it and where I added it right these are the customer and functions that stored in separate separate files and I'm using these functions through the code custom library ke bad we have to set the this is get resolved option this is again very important part so you remember so this parameters were shared pass to the Glue from the conversion to the Glue it is also

passed right so all these parameters are coming to Blue but blue will not pick them up unless and until you will specify the names which arguments to be taken as to be specified here as well you don't specify it in the code the Glue won't pick it up it won't identify them ok so whatever arguments are required from the JSON not all but whatever selective ID required I pass the names ok which arguments I need and use them further in my process Kasam constants have defined the constants now again the account Id comes into picture which I was telling so for focusing right now there is a change as I told you in the conversion as well there is a change in the columns the column types and the column names as well for both the vendors but majority of the code of the functionality was same so there was no point writing another job for it so what I did it is for the primary and secondary are kept the same job just what the things I use the condition so I created a proof Boolean parameter basically if the account Id equals to 9 so if account Id and getting is equals to 9 and accepted as string because this is another thing whatever values will receive in the Glue will be in the string format whether it's a number or any other field because the compared with the string is equal to 9 the value of Boolean true will be saved in the primary and using this book and primary as a identifier kind of thing so if it is the primary account these are the required columns and looking for in the code and this is the column mapping I am looking for sorry mapping to create in terms of ok this is the name which is coming from the vendor this is the data type of that field it will be a string because it is a csv file now this is the name I want it to take for my process my code my DB columns accordingly then I give this name and I specify the data type accordingly so if it's numeric column quantity for example specified is begin if it's a text column it's a string if decimal column then it's double specify the data type for that column and the name you will be using throughout the process and at the end of the database Global only focus on the specified here in the column mapping and along with this there are some catchy places like data studying table we have these columns and everything so this is the final names which we want to have in our data weight length width height lesbian ISBN underscore 10 ok page but if you see in this code I have written as I have been underscore 13 now this column does not exist but I have written similarly for if you see this pop date published UNESCO date not published weight underscore arms length underscore in these things because I am using this as to specify and in the conversions and all but I'll show you how are you primary is sending ESPN 13 only I can tell it is not sending so we have used the function has been conversion 10 generic function are created the specified as been underscore 13 this means that because it is sending eyes when 13 has been numbered and this column will automatically be passed when I am using further down 13 has a relevance similarly in the weight have added the unit in which unit it is

sharing the weight if it is by default storing in grams in grams then we directly right weight if other than grams then we specify the unit I think about the unit which unit we are receiving if it is mm that is what we are storing no need to write if any other unit we need to specify here and then the date format because that is also coming different for different vendors so in which format just sending the descending mdy format and date and then Year Without any happens or slashes ok thanks this is different for secondary condition and specify these things again again column map of secondary and Measurement second reading my father file name create a file path and I use the file path in the internal function of the Glue to read that file then I am getting the count and total records how much are processed in everything in logging it in the SMS service then specify the required columns are required and there is a quality check class function object in which I am passing the data and the required columns and the back end functions are taking care of it now this is the mapping function this is the input function and specifying the column mapper so which ever occur in it is primarily the primary column paper will be here if it is secondary and secondary column paper will be here and that accordingly the code will be back now you can see there are certain extra keys are there for example account Id vendor ID you know these fields are not coming in the field or no matter is sending that this information has to be added separately so in that case I have added a foreign key function in which I'll add these fields which I have taken in the argument and I'll pass this to the code added as a column so that when at the time of injection these fields are available filtration of bad records like if any eyes with 13 is wrong or the price of those filtration and no things are done here then after that filtration of valid is in passing the ice in column here what will happen if we have the weight column weight conversion no my system knows with column has an extension underscore thing so it will pick accordingly same for dimension measurements are passed it will take accordingly because underscore is there and when I'll show you the code you will understand better you conversion so the directly pass the haven't specified the data directly pass the thing because the system knows if I have written eyes been underscore 13 and I am passing calling the function has been conversion 10 it is it my system knows which column to be pink so again when I move to the library I will tell you again will get better understanding white so I am connecting to database separately to fetch that what you call it the country code from the White listing table we have whitelisting table as well and I am merging those country names if available for an ispn along with the available country code which are getting in the field so some vendors they send availability country codes in which which country the book is available so if we have other countries in the whitelisting table along with the available country we must share in the step and process further so like now in this case

here in the books in primary we were getting published date but in because in secondary we are not getting that so for date conversion function as specified this condition again if it is because in primary then only this bhajan step will be added as it will not add this step and directly process with the previous frame whatever it was here ok so this is the use the conditions and account Id in the same job to you know do multiple personality the same job with major to the code was sent then we add information like created by the blue Run Idea and adding to the created by field so that I can trace back if any error miss value is added so I can trace back which job Run ID was used to execute this job I can check logs for that particular job in the ID is true and secondly as requested by the panel itself some enters are not sending currency so the currency was coming as blank for so for those vendor know the currency to be directly add it here so we specify currency USD as Bugatti was not sending USB and we know it so we have had it in the code at the time of injection should be USD then I am calling this function insert into DB it is the mass of the original blue function I'll show you the libraries custom function inserting the data in the Staging table then I am closing the job and just logging the messages so this is how typical blue job we have not details we have certain configuration to pass on like the service role Spark versions everything the work type processor we are using number of workers so that concurrent network in a specified here so number of workers means how many cpu's or GPU is required for that particular job so if it's a heavy data job we specify more number of workers if it is ok ok data is specified less number of SO2 is the meaning of the advanced option we have these things so this is the job is ok this is the concurrency we were talking about how many concurrent jobs can specify it here temporary parts with is already specified the temperature changes or you know this is used by the globe we can do about it much at the connection the DB connection we create this is where we add this is the library path where I so yesterday I showed you this in S3 data insertion ATL fraud so this bucket is used to store code here we have add on zips add on lips five different Python files and put it as a single zip file in the HD bucket and specified this file path here is the library section from this file you need to get those Python files as libraries so yesterday if I understand what do we define those like how many types is decided automatically by the Listener Service application responsible to divide split the file at the time of intention when it is business service is picking the file from FTP it check the file size if the file size is more than 250 MB because it will get in Prime now so we have fixed the threshold so any file above 250 MB will automatically be split chunks and depending on the data in the file of the size of the file chance will automatically be created so that is predefined at the time of business service after the chunks are already created the entry is created in the feedback table as I showed you here this feedback table here the

entry will be created individually for each time and according to this we have a separate Run ID for each junk and that Run It will be used in the step function individually I have a separate country like this so this is the run ID this is the vendor name and this is the time stamp so will have processing and they are 26 different things to wear 26 different Run ideas of India with the close time stamp because they are already getting process right out of my scope that is with the Listener Service media I am showing you this is the concurrent joke how many concurrent processes can happen have showed you in the system conflict table right concurrent job thing if it is specified to the only two concave jobs I am allowing like for Instagram it is 25 at specified concurrent jobs to 25 different jobs can run parallel so this number and specifying here in maximum concurrency how many maximum concurrent job I am allowing is cross it will give an error the maximum conferencing is accident that way please try let me show you the libraries folder so this is the add on lips we have 1 2 3 4 5 6 7 files we have one is the constant file kept this file because I define the constant which are using throughout all the different different files and this is environment credentials ok so so might be different for debt this might be different as and here are specified the connection name all the information I am using the DB name the blue DB name the field master table white list table the table names which are there the field you know writing the logs right so that path that is also environment specific database name Staging table name error table name this information are constant I am using other than that they have different different files to different processes write one disease for extra data so remember I told you about that white list tables if we are getting some country names if available so you doing that thing when creating a connection with the database we are fetching we are wearing the white list table for particular ISP and checking if they have the country names if yes then we are merging them with the available country column which is coming from the field and then updating the complete data set so that functionality is here similarity date and so in this library what are doing this we are doing this date thing so that published date we are getting so we have different formats as specified some of them we are getting so if it is this format if you remember it was made mdy so it will go to mdy split the date accordingly and then returned in the format DB can expect or which is required by the bandwidth so year month then date standardized the date which is coming in different different these are the different different formats in which different vendors are sending the date and we are standardizing it to why MD with that function is written here so if you see there are two functions this is the actual logic function and this is the group function because in blue we have data frame collections from Data frame collection we extract the data frame and then on data frame we applied transformation on a particular column so I was

extracted the data frame from the data frame collection on a particular column are called this function this function will be performed on a particular column that's why I have to have two different functions in the group we will call this function but actual logic function is two things like this you will face in glue where you have a separate functions like the standard function which will get context from the Glue the data frame collection from the Glue and then extract the data frame out of it and then call the logical functions over a single column like we use Lambda dynamic function or our specific function so it is close to biosphere this is another separate function we operate foreign ke edition so the foreign keys I told you about the vendor ID the time which is right now the United States default and account Id the source country so this information we were getting in the JSON but not in the field specify this information added to the data set this file this file deals with all the types of has been conversions we are having from eyes with 13 to ISBN 10 from ISBN 10 to ISBN 13 or some vendors they send mixed conversions like this and I will 13 value and has been 10 value in the same column so for them we have a separate function has been conversion as well which will check this has been is has been 10 or 13 if it is a 13 it will convert into 10 and save the value in 10 column separately if it is a 10 value it will converted to 13 and save it into teen column to be picked to be added to ISBN 10 column and then renaming this column to ISBN now this is what we have in our Staging table is called rename has been underscore 13 column here in this function to ispn so that the mapping should not break while inserting the records because if I keep it has been 13 the system will not know show all the ice cream conversion parts are in the has been conversion file similarly for measurement of weight are specified so if you remember it was great underscore arms right so if it is ounce my food knows which unit which values to be pink if it's found what to be picked and if it's and need to multiply to convert into grams and then rounding it off to 2 units returning the greatest grams right then splitting it this is the again the logical function this is the blue function so blue function and identifying weight splitting it getting the unit converting it and renaming the column two weight because that is what we have in the DB wait no underscore right similarly damage in conversion so as specified in length underscore in right so sports note if it is in I need to take this value multiply and convert it to millimetres mm and then store it in the function when the data set and then further in the TV then we have this function so in this I am checking the data for validation like you know required columns are there or not data cleaning vice president is there in the column names because sometimes the vendor sending column name with a white space then again my system might fail so for that there is clean up of column currency value is also there so if a currency which is not there in the system that record will be treated as an error report chords I get the

reason why this record is in error I have a reason for that records like any of them is so which column was none that also we need to specify required is also there in this file only safest is a very nice in 13 or not if it is very nice in 10 or not those functions are here defined in this file this library then we have the utility function by other than this if anything is left we are dealing with that years on the same miscellaneous file like one printer is sending 3 alphabet country code and for others we have to help so for that purpose things and cleaning of things are done in this file that to lock whatever locks I am writing so this is the lock logic so this is how I have created the JSON and Aryan and what not which is required so I'm just calling this function there and the remaining code is written in this file this is the logic for the logging logging errors file so when the file is process success told you that file is moved to process feed buckets right so this is the function which is responsible for that successful it will move this file to from the processing bucket to the process bucket so if there are unnecessary null columns are there so we drop them restricted countries we need to fetch so we set the distributed countries and update it in the code sale restriction column we have and then insert into db so this is the whole logic to insert data into DB this is my mass function I am using the table name so and the columns I am updating the scheme of the data frame based on the Schema in the database so that they should not be a mismatch of the data type while inserting the data in the table and then finally this is the which is directly calling the catalogue table from catalogue as you can see so this is not referring to the baby of the blue data catalogue database it will insert there and from there it will be sync with the RDS so these are the libraries which we are using the folder and we uploaded it in bucket here jobs we can directly you know edit it in the editor itself and save it or in the bucket in the ETL jobs in the scripts if we upload the same name while that will replace the existing file both ways we can do for Lambda directly go to editor and change the code there is no file storage or floating libraries which we are using to conversion layer we have this conversion library is required for the additional which are not by default there in Lambda three of them we need to install them and then we upload that Sadi Nahin to be changed but yeah this keeps on changing I don't have any code mappers maps Mein we are using csv files result tell you and again explain the XML conversion code with the blue process we can move back this how we have different different jobs created for different different files and I think this is another I like to show generic Phoenix job ok so earlier we had this separate conversions and separate column things for separate XML vendor but as the XML vendors group like seeing more similar structure and those things so I created one job for the onex vendors because they are similar in everything so in the column mapping is specified I have written here as well in the comments if you ever 13 is there in what values of the column mapping is same for

all the XML vendors so I created a generic job didn't disturb the existing Windows but for the later on additions when were there so I added this single job any single changes required to access the account Id again the same thing ideas are there so if there is one change for scholastic Lewis one was for the sale UK if the account Id is created a Boolean variable used basically change that particular thing if it is not scholastic then discount to be added if scholarship us is there discount should not be picked so that is why this is separating like if then UK is there we need to pick the quantity from the XML if not you are not taking quantity from the XML so like this we can add more condition if required for a particular mapping this condition Kunal is the same that I asked yesterday like if there is some more checks or criterias needed to be added while the fields are process want to remove a column ride a particular column so this is the column mapping where we done any column which is not mentioned in the column mapping will not be taking so whatever columns to pick we have to specify and the column mapping and then the names to be proper with the database and the data type accordingly write data type has to be there but the columns are coming differently from different vendors right so this is generic for the OLX OLX XML files we are getting converting them I have standardized the names of the tags in which app store in the value so I am pretty sure what name I will be getting in the XML file that is why I create a generic only for the earnings for other vendors we have a separate job for separate file type because if you see in the book are seen in primary secondary or not saying the column names are different so whatever name is coming in the file I am using it and then renaming it according to my convenience or the systems convenience in the oven is these names you can see in the first column these are given by me at the conversion level I'll again go back and I'll show you the XML conversion thing again you will get a better clarity there has been so these names are specified by means I know which name I'll be getting for which value so that is my created a generic job for this one because I know which columns I need to pick and that to be picked from all the XML files could be possible like for any particular comics file or something any any of the column is missed then also This will work right it's not what will happen if any column will be missed so it will be added in the data set by my conversion code but the column will be black suppose certain tags are not sent by a particular vendor and certain text it is very common this is the biggest so what I have done is if I find a tag I picked the value if I don't find the tax I keep the column name but I put it as blank or not I have all the columns so if they supply the value that value is take if you don't supply the value itself so it will be null again if you want to not to update a particular field or a particular column for a particular what you can do is if I remove for example I don't want title to be updated person right I remove the title from the column number now what will

happen in the feet stitching and secondary considered because it has less than 13 after converting and title on order quantity price of the publisher when like I think 829 columns right but in the Staging we have way more columns and that so when the blue will insert these nine columns data all the remaining columns which values are not supply by the clue for that particular job will be updated as by default null ok so if you don't want to update the title here you will remove the title from the mapping title will not be there in the data set so when it will be inserted in the DB in the Staging table the title will be done after the blue inserted into Staging table now this further process I am talking now we are through with the blue part blue successful now the choice choices can do I want a trigger migration or not so what is the trigger migration what is this migration all about so basically we have to store procedures created one is the synchronised bibliography and what is the price in availability price and availability I am using an upset command is it will insert an ISBN if any conflict is raised which conflict the combination of ISBN and account it already exist in the system in that case it will update the fields instead of inserting it is the combination of ISBN and is not there in price in available table it will directly insert the record it if it is already there conflict raises then on that conflict and updating that feeling so updating field there are certain values which Panvel team wanted me to override every time whether whatever they are coming but the remaining things we are not updating it as checking if the value if it is Null we are not updating if it price and availability table so these are the columns which are available in the price and availability table and we update them I think we update every time because that is the information they don't want me to sustain bibliography it will try to insert the SBM if I have been already in the system and it raises the conflict I am updating the has been values so here I am checking all the fields if it is Null I am not updating if it is not null and updating it so these are the columns in the Bibliography table which are being updated and this is done based on the priority of the priority concept based on the priority only we are updating the values so I just want to confirm it like for a particular is being there will be only one entry will can be repeated multiple times but the combination of ISBN and account Id is unique because there is a possibility 1 ISP is shared by Ingram account along with Bugatti and the population is also sending the same so we have three account sending the information but with different prices are currencies may be having a different price in a different discount because in must be having a different price in a different discount in gram must be having different price in different discount for the same issue so that is why in the price and availability table we have a combination of isbm account we can say like for a particular account will be there if there one has been with shared by four different accounts we could have four different countries it is combination base in Bibliography it

is just the isbm unique is being decided in the Bibliography information like title if we share a title that won't very based on different different accounts if income is sending because in ascending and hypochlorination sending the same is given the title will still remain the same so as many different other fields right so that information is stored in the bibliography and that is unique for the ISBT in the price and availability and discounts that change from account to account so that information is kept separate in the combination of ISBN and account these two stored procedures were managing feedback as well if you see status on the migration status so these two columns are updated accordingly so when the migration is completed migration means the p and synchronisation is done it is completed when the people of synchronisation it is market Bibliography status as complete when both are completed the entries marked as completed in progress ok procedures are run at the time of separately on a schedule basis because the data is used and the data is interesting continuously so if we don't what will happen if it will lock the table is happening lot of you know migration updation is happening then we are wearing the generating quotations as well to read write operation continuously will happen on the table so it will lock the tables so for to remove that thing what we did is which I have put a schedule so it is capital expenditure initially later status for it like this pending so it will pick the pending one run the store procedure and update the value in once it is completed it will mark as completed so it is running separately for P and separately for Windows 10 minutes every 10 minutes it is taking more assuming 10 minutes is required to update data plus minus and what happens if this process is running and others want to read write or do some quotation or whatever updates on the front end then what happens that is working fine after we have introduced this updation thing for the incoming this is done only for the incoming ok so because incoming feed the frequency and data both is used process it will take an update accordingly using this process the load is not that much because system is capable of handling that much of load but you know 12 or maybe three two operations at a time along with the quotation generation and everything it is fine but 5060 updations at a time along with protection was locking the table two or three jobs are writing the same table if I schedule this is helping me right mainly is because if this job is running so next job should wait before writing the same table the other the report generation and all they are reading the table but the migration the PNA synthesization is writing the table so jao if simultaneously try to write the same table or the same is welcome time to avoid that thing it is scheduled at one time only one migration should run and this two can run parallel because another table and figure is another table but on the same table should not run parallel one by one so that they should not lock to updating Simons simultaneously so this is where this choice comes in picture ok for the

incoming field we have specified KV don't want immediate migration data can wait and can be updated with a certain delay right but if it is a manual upload Panvel team must have uploaded how many thousand 5000 10000 maximum upload in that case we specify the trigger migration should happen immediately be specified this as true so after the Glue This trigger magnetic step also happens and when this is completed it is marked as completely immediately and the data can be the data get reflected in the system immediate basis so this is the manual feed right so anything uploaded by the manual feed the data will directly being reflected after this job is complete the data is reflected in the system in advance search and quotation wherever they want to use but for the incoming feed we have a certain delay so you know because the frequency is too much approximately I think we have 15200 injection study Jo sleeping hours to avoid the delay because the converting wanted to see the data immediately after the process so be added to check here in the choice so if it's an incoming feed it is false it will wait for 1 minute in one second and it will end but if it is a manual upload it will check it will be true it will the migration and then it will complete this was the process for the migration thing anyone anything I am ok to location from myself just tell me what what if I want to add one more check over here some place check say any any example we can take like if you say in the manual process I want to update directly for one particular vendor and not for the others or something like that of onix for that matter so anywhere if I want to add any more should be the right place to do that in terms of migration you are talking about triggered immediately or not ok so first thing you need to add condition in the Java code problem in the Listener Service which is creating this JSON and sending it the parameter should come here along with the value then once you have that value you can add a choice or a condition there and you can update your diagram accordingly so this is the edit state machine thing you can click on here and you will get the control of the company we can design a drag and drop you can drag drop it so these are the service all services available here you can drag and drop any one of them and place it wherever you feel like and then configure it accordingly so this choice if I check this is the configuration of choice we have set the value of set the rule what condition if it is true then what Shubh benefits false what needs to be done you know and then you can give a direction to get it by default What to do and if it's true then what needs to be done if else and if you want to change something with respect to code to then you can directly click on the code part you will get the complete code of the choice like this execution migration needed to this exhibition in migration in JSON field and execution input is the input parameters what we have passed so to access the input parameter we have passed to specific like this tell him input migration this is to specify the input and this is for the particular field we

need to update the value you get that you can write a condition of your own similarly you can service here as well if I could design again like how blue is added to you click on the Glue will see like how the globe is added so these are the argument passed for the Glue ETL job name we are fetching from there an ID vendor ID you know all this information from their passing it to Blue wait for task to complete so that it should not proceed further wait for the response from the Glue and then next state so that the arrows are formed and modify this flow for the step function and what about the Java code for the same you going to take us through that developer we have different business services so we are clear with the process sofa questions any doubts anything which is not clear with respect to step function Lambda probably blue database with respect to the injection part yes whatever you explain seems good so you would be better knowing what if you have skip anything but happy birthday wishing because right now it seems fine

Recording: 4 Stepfn_ETL_GLUE_Python_code_KT-2.txt

was all that was done in Python in India Chaudhari respect to the functional flow configuration part you know which area target for which Service I think right so again the same thing that is go through the things ok in which we choose the account Koderma separate template for it which have all the columns in whichever columns are filled by the team and uploaded only that gets updated are there will remain null as the Staging it is not right we don't have the option to choose the account Id right so in account Id what will happen is without the account Id in the case of value of load if you see this execution may we have the speed time PNA right this will come is bibliography and when this will come is Deep Blue the trigger migration for PNA table should not have the price won't be there the currency of these are the crucial parameters with other information is there so only migration will happen in that case if the field type is video and that is selected from here just check for this disable overriding for send Bibliography disable override is done then what will happen is remember the priorities we have in the configure Masters with the vendors why I will show you where is the mental Masters this is the priority right the vendor priority this priority value is being used in the Bibliography table to decide which vendor information should be trusted more so we have started this priority number from onwards one we have reserve for this checkbox if we normally upload Bibliography table directly bibliographic sink will happen the process will be normal no issues if we check this box disable overhead then the Bibliography upload will happen with the priority one 209 any other critical overwrite that what is over what is updated by two vendors are having the same priority then what happens is there a conflict or something is not that it will update you

can update two Cannot update one all the vendors starting from 2 onwards so that one should not be overridden by any other number one ampere preserved for the this check boxing in the Bibliography yesterday we discussed about the vendor priority and the priorities for Windows 10 publishers and 325 for Instagram food distributors in 74 Indian publishers and distributors so currently you said one is reserved for this disable override so how which one would be followed the vendor one priority or the disable over right now account in the publisher starts from 2 then the first distributor I think privatisation Sangram distributor is 4 OK 6 for some other which printer is first then these three distributors then the other distributors then the so the disability so the disable override checkbox if we enable it it will not allow anyone send the priority is one this is the whole injection process should listen that is the Listener Service I think probably the next session is of business services increase the feedback entry and then that only triggers the step function so once the step function is triggered that is where my partner is provided to me then from there on the conversion you know the crawler probably and the blue and the trigger migration that comes under my territories or questions with respect to the the logic or any particular thing is bugging you please how do you tested locally like Listener Service can be set up locally so in sync with the details of so how do you guys tested in the local environment and I was developing cricket test job there itself on the environment only that job with the minimum configuration possible so I guess that setting locally is very difficult and especially with the happiest security construction install certain libraries in a little district and the tape and environment I need to check I write the code the console prints and everything the checkpoints wherever I like to check what data I am getting in which step and just not the whole code I don't want whatever want steps I want to check at a specified those only and ended there and in the console I can see the printing and I can see the data and whatever it is giving you work [Unintelligible audio] and I think so we are good with Lambda S3 step functions RDS and group these are the main services which I am interacting with other than that the two services which I am interacting with the cloudwatch with the locks are available and this is one more service simple notification service SMS this service we are using for the login to transport the message notifications are there this is the one page is interacting with the blue the feed processing law for this channel we have created a published subscription is published and it is subscribed by the other whatever it is there open this one so there are two subscriber to it one is the whatever this API they have Java people and the other is one my email ID so that I can personally monitor with vlogs if anything I need to press back or check the two subscribers every time the different format which step is completed what is the count and how much information we are sharing from the job if it is filled

what is the error at which step it failed and that what was the total count processed recording whatever information is there that is supplied in this is the channel by using for that pretty much distance some part of the functions would be covered in the business services which part triggering of the step function because this entry of a lot of entries created these entries so these are created by the Listener Service responsibility get system conflict about the job the path and then based on the account of initialise this particular function instance and then the process happens we get the input resistance and the first deployment se concerning the goods are directing available for glue and Lambda so direct Court there that start library cards update for this extra files we have so that we need to update this I want to introduce so what is happening is in this conversion XML conversion codes are sent like for a particular value content audience time may be the code standard 0 2 [Unintelligible audio] used OLX decoding is there in this subject this tag is coming in the XML subject schema identifier or a short tag is 2067 so this tag will contain a code for example 03 not this 03 has a significance for this list they have shared here from value whatever code they have in the system what is the meaning of that code is mentioned here 03 is coming the it means LC classification right to is coming at a subject category if one one is coming It's a bisect reason for the numbers and their respective meaning to it so the number 11 or 12 in the subject category aur subject schema identifier what kind of subject code it is this will be used information right so this csv to see here based on the tag name I have created the course I have copied this all so there one is the value second is the description these are there so at the time of conversion when I am dealing with this particular tag or particular column I replace the values effective meaning which is stored in the case where it from there I use this course to map and replace the description and then I insert the description instead of the code in the data that is where this code csv are coming to picture like for the XML files initially we were storing the starting tags and dash in the TV so that is to read the file that is different that is for the listener to split basically that tags are specified to split the file because if some XML files or even larger than 250 MB so that specify that is in the field master in that was because suppose that's more than larger XML files so in that case developer needs to split they don't know from where to which tag to start in which tag to end to split the file so for their help in utility in the feed master table we specify the tags so the Java code knows from which tag it should start in which state it should end in order to split the file in the correct format is the format will be wrong in that will give to error anyways these are the codes basically this for this is one example a subject schema identifier code right so Ek subject kab hoga [Unintelligible audio] what is tension so this is audience range qualifier so we have certain values again and we have a description

similarly audience range so this is how to store the codes so the value and the description choose exactly here on this page values and description now when I get the code I get like this 11 15 17 so while after insert before completely converting in the file and passing it to the ATL layer and replace the code 11 with the specific value which is the US School grade range in the column whichever I am passing this tag into I will pass your school grade range and that value will pass through the ATL and further to the database certainly no relevant meaning what is the value what kind of code it is who is this time code value noise that meaning to the other tag second tag which is the next step which is the complete example like this if you see this one so it is up to us gold rate 4 to qualify values 11 that means us whole grain 04 has a pressure value and range values for the list to it decision value they have another list if it is 04 that is 2 in the ETL jobs do you have the exact convention Sudheer I am dealing with the tags and converting them into csv in the ETL layer I am only passing the csv file ok so this part is handled in the lamp light conversion exactly so the conversion the handle in the Lambda itself Gramin checking the values of the tax standardizing it into csv and then passing this csv to the ATL market the XML owners we are getting it is very complex and very nested directly would be able to handle this next level of which is very all the always available right this phone is the responsible to fetch the values from the tags converted to a csv and then share it to eating from this files mixi have when the environment variable types I am using this path to fetch the files of the csv files read the codes map the codes whatever quotes I have received in the field and replace it with the description relevant relevant description is done the csv file is created with all the respective columns and their values with structured and then it is passed to the ATM for the further injection [Unintelligible audio] writing so yeah I think that's all for myself with respect to injection any aspect of induction you feel you don't understand or if you have some kind doubts please try to ask now or properly later as well innovations now just love you good or that's all romantic part these are the two main things manual upload from the FTP we start the conversion be transformed into initial and inserting the database and then we have the migration for which we have two stroke procedures that is done we can conclude the call hi sorry sorry thank you so much do let us know if possible could you share any document of the same all that you have explained in the last two days because you also know it's very explore things thank you so much thank you thank you thank you thank you