We get 3 types of tickets:

1. Etl jobs (mostly)
2. Web applications
3. Data science related ML modelling steps

Hive to snowflake migration is also happening

Whatever change we do we will be doing in the target project (do a **git pull** and fetch the latest code) that only inside the infra folder of the target project

Jenkinsfile.infra pipeline -> for spark on Kubernetes, we need **docker image**, to store that we need ECR repository, Jenkins.infra pipeline will create an ECR repository, we will run terraform commands (code related to it is stored in terraform/setups folder) **which are used to create ECR repository** -> which will help In storing **spark on Kubernetes image**

**Jenkinsfile.infra** is used **to create ECR repository to store Spark on Kubernetes (afaik and other images as well)**

Jenkinsfile.spark\_submit -> we will be building the image for sand box, using that sandbox image to build preprod image, using the preprod image we will build prod image

Bandit code quality checks will be done, if failed, pipeline execution stops, else pipeline execution continues

Creating docker image is a sandbox related stuff and will be done in Jenkinsfile.spark\_submit

2 images are to be built

1. Base image
2. App image

Base image is built when it is getting built for the first time or if there is change in requirements.txt or in docker base image, or if you explicitly force to build image

Jenkinsfile.dockerpull -> this is needed for preprod and prod, as this file is used to pull docker image from previous environment (order of environments sandbox->pre prod -> prod)

For preprod, we will pull image from sandbox, for prod, we will pull image from preprod

Docker image is only pulled in preprod and prod and that is because we create docker image in sandbox

**to perform pulling image from previous environment and tagging it in current environment we use jenkinsfile.dockerpull, using this we can also push the image**

for required codes in src, we need to have pipelines for required ipynb files, for each ipynb file based on requirement we need to write Jenkinsfile.infra, Jenkinsfile.spark\_submit, Jenkinsfile.dockerpull, there might be an pipeline existing already in EMR way, but we need to change that to **spark on Kubernetes** using the above 3 jenkinsfiles based on need

# changes to be made for ETL:

You need to change business line and project in infra/terraform/setup/env/common.tfvars, if not sure of business line, need to contact ticket reporter

## In spark\_submit:

* Change RECIPIENTS\_LIST, CREDENTIALS\_NAME, ECR\_REPO\_NAME (you have to check the name in the ECR.tf or the tf file where the ECR is getting created and change it accordingly In Jenkinsfile.spark\_submit and Jenkinsfile.dockerpull)
* TASK\_NAME and FILE\_NAME should have the extra information that you give while naming the spark\_submit file FILE\_NAME is used for echoing it on jenkins console output, FILE\_NAME/TASK\_NAME should have words separate by underscore, it should depict the code that it is executing, if you have multiple ipynb/py files, then most probably they will have proper names, else case if it is not having proper names you need to ask/understand and give names accordingly, if it is a single file, then most probably giving project name will suffice, the task name is **used for naming the pod in Kubernetes cluster, where the jobs run**
* FILE\_PATH-> it should be the spark file path, i.e., folder of the py file(the path where the py files are stored, to decide this we have to look at docker app image, so that we can understand how and where files are stored (in our local, if we have src in root folder and main.py is the py file and in docker app file, you copy everything and for executing spark-submit command, you can just give .src/main.py to assign it to filepath), if ipynb files are there, then actually ipynb file that is converted to py file, command is stored as make command in makefile and ipynb files converted to py are stored in specified directory in the make command) + name of the file.py
  + Ex: build/<ipynb file to be deployed as spark job>.py
* SPARK\_VERSION= “3.1.2” 0R “3.2.0” (any one of these based on the version used by requestor), **similarly** change the same spark image version in the dockerfile.base and add the above version as tag for spark-executor-base:
* I think I forgot to include project\_name in backend-config files for environments in terraform
* Set ECR repository name in docker jenkinsfile.dockerpull
* We need to pay attention to how files are getting copied (**in to what directory**) and which files are getting copied, how are they being used, for example if you are importing a package from other folder in to src, then importing them must be viable with the file structure in the docker image structure as well
* If you have have ipynb files/py files with spaces, we need to remove them
  + For py files directly
  + For ipynb files, you need to change the name of the task and name of the file, so that when ipynb files are converted using nbconvert (if you have spaces then a single file is treated as two different files and will lead to errors), they will have proper naming structure

docker pull runs only in preprod and prod

infra pipeline -> used to create ECR repository

check about bandit

dockerimage is created in sandbox

# Questions to ask:

Where is prefix of terraform code being used?

File\_path for files which don’t have ipynb files? Tried to check using jobname, but no ipynb file for jobname as well

Mpm\_pricing has 2 files, 7B and 7C, I gave 7B

# My understandings

* Deps is command in makefile to install dependencies in Kubernetes cluster environment
* Whenever you find a connection (flow of control for commands) between 2 files, then take a moment and register it in brain