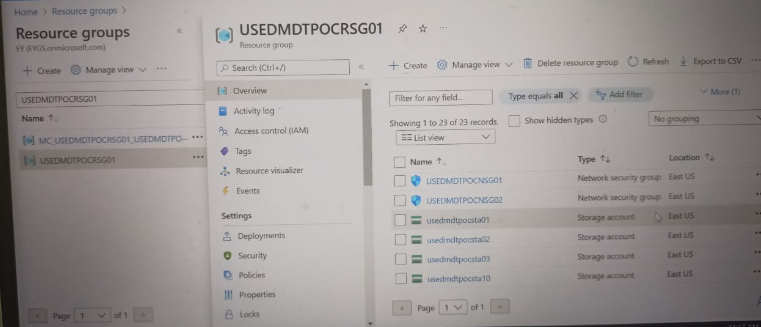
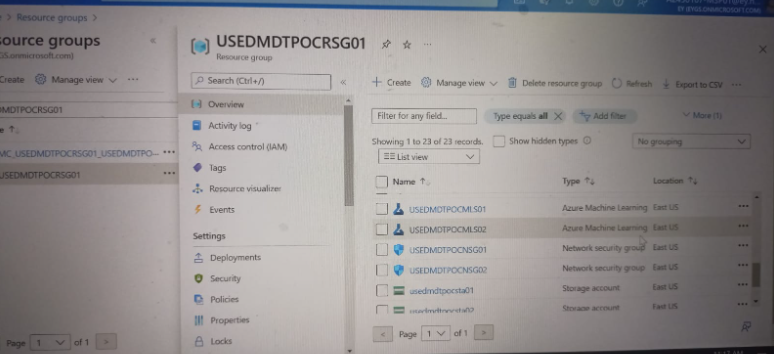
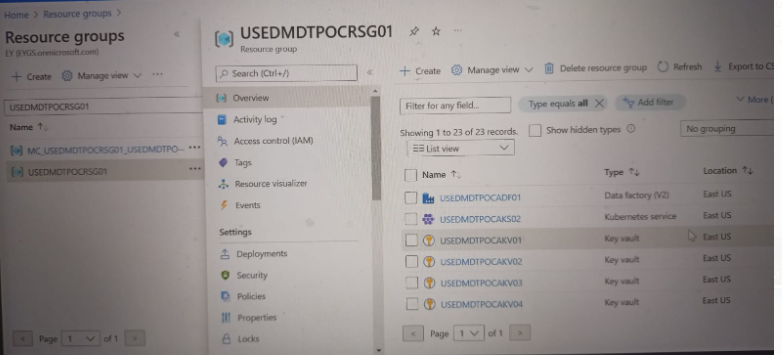
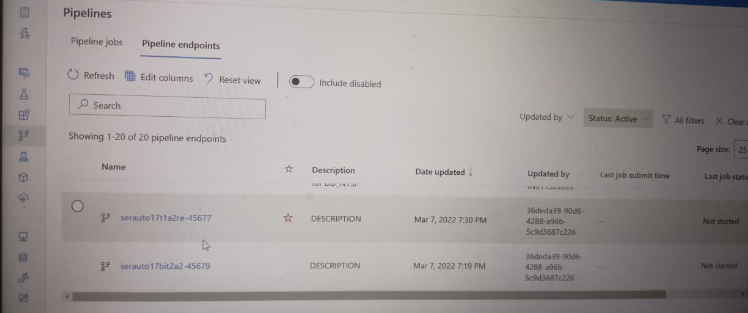
MLops on-boarding modules creates endpoints for Real time , for Batch inference Pipeline ends points will be created on portal

Real Time Models – without storage – Type1 (CPU -0.1-0.3 & Memory – 0.1-0.8GB), Type2(CPU -0.1-2 & Memory – 0.1-8GB)

Batch inference Models – without storage –Type-1 (use for azure devops to store the model)

Batch inference Models – without storage –Type-2 (use for storage account to store the model)



Batch inference -> end-points

Real-time -endpoints