Data commons:

Steps Involved:

1. Downloaded files from dbGaP application ( I have downloaded all 11 studies from dbGaP)
   1. We need to use Aspera Connect to download files from dbGaP
   2. After downloading the files, download NCBI SRA tool kit and configure it using dbGaP repository key and the command has to be issued under the dbGap workspace (get the dbGap repository key from PI). Save the key as .ngc file
   3. A dbGaP project directory should be created as soon as the dbGaP repository key is imported.
   4. We need to configure the SRA tool kit. For this go to bin sub directory of the tool kit and run ./vdb-config –i

(Use this link for better understanding <https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=toolkit_doc&f=std#s-4> )

* 1. After configuring the SRA tool kit use vdb-decrypt to decrypt the downloaded data

1. After decryption, Parse the data dictionary XML files using (data\_dict.py) script and output is in JSON format
2. Load all the JSON data dictionary files to BigQuery (loaded all the files into single table)
3. Shortlisted the variables- pete
4. Currently, working on getting data for the variables shortlisted\*\*
5. Data to be loaded to gen3 windmill data model\*\*