1. Add your chosen data to your GitHub repo (I'd suggest creating one that will be used to load data into your Cloudera VM)

2. Clone (or pull) your data into the Cloudera VM

**cd Documents**

**git clone https://github.com/Gayathri-Shanmuga-Sundaram/HDFS\_Hive**

3. In the VM, use the terminal to create a HDFS folder called data16 located within the /user/cloudera folder.

**hdfs dfs -mkdir /user/cloudera/data16**

4. Copy the data from the local filesystem into your HDFS /user/cloudera/data16 folder

**hdfs dfs -put /home/cloudera/Documents/HDFS\_Hive/HR-Employee-Attrition.csv /user/cloudera/data16/**

5. Using HUE, create a new database called data16 and load your data into this database as a hive table.

**hdfs dfs -chmod -R 777 /user/cloudera/data16**

**hdfs dfs -ls /user/cloudera/data16**

6. Come up with a question that you want to be answered using the data - Create an SQL query to answer this question (do more than select \* -- demonstrate filtering and aggregation (such as count or average)

**SELECT department, AVG(monthlyincome) AS avg\_salary, COUNT(\*) AS num\_employees**

**FROM hremployeeattrition**

**WHERE age > 30 AND attrition = 'Yes'**

**GROUP BY department**

**ORDER BY avg\_salary DESC;**