DASK AND SPARK ASSIGNMENT: HEALTHCARE DATA

INTRODUCTION

For this assignment, we will be using some simulated electronic health records (these are not real data!). This is a common sort of dataset for health care systems to use when tracking all the patients and the outpatient activity. You should take a few minutes to review the datasets using Excel, read the descriptions, and understand how they fit together. We will only use a few datasets in this exercise, but you could explore all of them on your own.

- OutpatientVisit.csv
  - This is a table with a row for every outpatient visit. This file can be linked to the patient table using Patient ID, the clinic table using ClinicID, the ICD table using the ICD_1-3 columns, and the staff table using StaffID.
- Patient.csv
  - This file has one row for every patient, and includes demographic information about each patient. This file can be linked to the outpatient visit file using Patient ID.
- Clinic.csv
  - This is a table of all the clinic IDs, and if the clinic is primary care, specialty care, or emergency department. There should be one row per clinic ID. You can link this file to the OutpatientVisit file by clinic code to see if a visit occurred in a primary care clinic, a specialty clinic, or the emergency department.
- Staff.csv
  - This is a table with information about each staff member. This can be linked to the visit file using staffID
- Mortality.csv
  - This is a table of date of death for each patient (if they have died). Patients who have not died are not listed in this file. You can link this to the other files using PatientID
- ICDCodes.csv
  - This is a table with the description for each ICD code. ICD stands for 'international classification of disease'. Each outpatient visit is associated with 1-3 of these codes, indicating what diseases the patient has that were addressed during the visit.
- DiseaseMap.csv

- This is a table that maps different ICD codes to clinically meaningful categories. For example, there are many different diagnoses for diabetes, but we might wish to generally call all of them 'diabetes'.

QUESTIONS

You will use these datasets to answer some questions listed below. Please implement a solution in Dask, and then also implement a solution in Apache Spark using the SQL style commands I went through in lecture.

1) Use Python and Dask to answer this question. I am interested in studying depression in our patient population. I would like you (the data scientist) to answer the following questions for me (a non-statistical leader):
    a. How common is depression in our patient population?
        i. This could be implemented a few different ways. You could simply consider everyone with at least one visit outpatient visit 'at-risk' in the denominator, then calculate if they have ever been diagnosed with depression. This is the simplest approach and fine to use for this assignment.
    b. Are depressed patients more likely to die than non-depressed patients?
        i. Again, this could be implemented a few different ways. You could consider the outcome to be mortality at any point and ignore differential follow-up time for this assignment.

2) Repeat number 1, but this time use Apache Spark to implement your solution.

