Hive Project

About data

When you browse to the techfield_bigdata/data/mortality, there are two CSV files which will be the input data in this assignment, `events.csv` and `mortality.csv`

The data provided in events.csv are event sequences. Each line of this file consists of a tuple with the format (patient id, event id, event description, timestamp, value).

For example,

```
1053,DIAG319049 ,Acute respiratory failure ,2924-10-08,1.0
1053,DIAG197320 ,Acute renal failure syndrome ,2924-10-08,1.0
1053,DRUG19122121 ,Insulin ,2924-10-08,1.0
1053,DRUG19122121 ,Insulin ,2924-10-11,1.0
1053,LAB3026361,Erythrocytes in Blood,2924-10-08,3.000
1053,LAB3026361,Erythrocytes in Blood,2924-10-08,3.690
1053,LAB3026361,Erythrocytes in Blood,2924-10-09,3.240
1053,LAB3026361,Erythrocytes in Blood,2924-10-10,3.470
```

- patient id: Identifies the patients in order to differentiate them from others. For example, the patient in the example above has patient id 1053.
- event id: Encodes all the clinical events that a patient has had. For example, DRUG19122121 means that a drug with RxNorm code 19122121 was prescribed to the patient, DIAG319049 means the patient was diagnosed with a disease with SNOMED code 319049, and LAB3026361 means that a laboratory test with LOINC code 3026361 was performed on the patient.
- event description: Shows the text description of the event. For example, DIAG319049 is the code for Acute respiratory failure, and DRUG19122121 is the code for Insulin.
- timestamp: Indicates the date at which the event happened. Here the timestamp is not a real date but a shifted date to protect the privacy of patients.
- value: Contains the value associated to an event. See Table 1 for the detailed descrip- tion.

  The data provided in mortality events.csv contains the patient ids of only the deceased people. They are in the form of a tuple with the format (patient id, timestamp, label)

Descriptive Statistics

Computing descriptive statistics on the data helps in developing predictive models. In this section, you need to write HIVE code that computes various metrics on the data. A skeleton code is provided as a starting point.

The definition of terms used in the result table are described below:

- Event Count: Number of events recorded for a given patient. Note that every line in the input file is an event.
- Encounter Count: Count of unique dates on which a given patient visited the ICU.
- Record Length: Duration (in number of days) between first event and last event for a given patient.
- Common Diagnosis: 5 most frequently occurring disease.
- Common Laboratory Test: 5 most frequently conducted test.
- Common Medication: 5 most frequently prescribed medications.

While counting common diagnoses, lab tests and medications, count all the occurrences of the codes. e.g. if one patient has the same code 3 times, the total count on that code should include all 3. Furthermore, the count is not per patient but per code.

a. Complete techfield_bigdata/hive/event statistics.hql for computing statistics required in the question. Please be aware that you are not allowed to change the filename.

b. Use events.csv and mortality.csv provided in data as input and fill Table 2 with actual values. We only need the top 5 codes for common diagnoses, labs and medications.

Metric	Deceased patients	Alive patients
Event Count		
1. Average Event Count		
2. Max Event Count		
3. Min Event Count		
Encounter Count		
1. Average Encounter Count		
2. Max Encounter Count		
3. Min Encounter Count		
Record Length		
1. Average Record Length		
2. Median Record Length		
3. Max Record Length		
4. Min Record Length		
Common Diagnosis		
Common Laboratory Test		
Common Medication		