Homework #3 (SQL and XML)

Due: 11:59pm, Friday, March 10, 2023

Points: 100

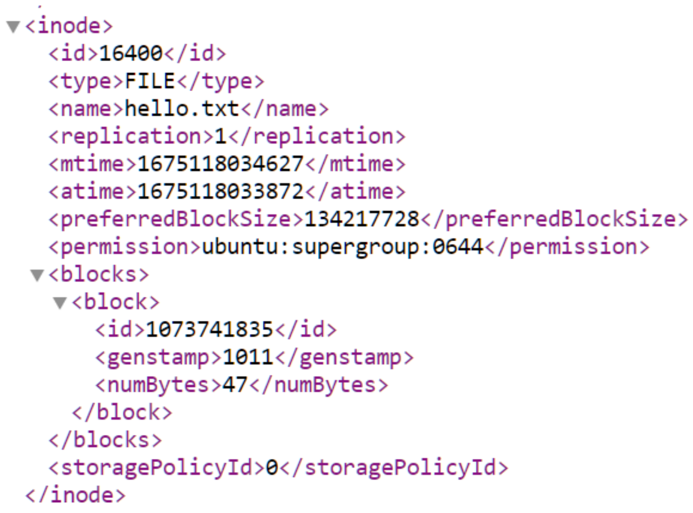
In this homework, we will take a fsimage file and store its inode and inode directory information (as illustrated below) in MySQL and write SQL queries to retrieve or analyze the content. Remember that XML contents are case-sensitive and need to be stored in MySQL databases as such too. For example, “FILE” needs to be stored as “FILE” instead of “file”, “DIRECTORY” needs to be stored as “DIRECTORY” instead of “directory”.

The homework assumes that you have created a database dsci551 on your EC2 MySQL, a user dsci551@localhost (with password Dsci-551), and grant all privileges on dsci551.* (i.e., objects in dsci551) to the user. You can log in as root and execute the following to satisfy the assumption:

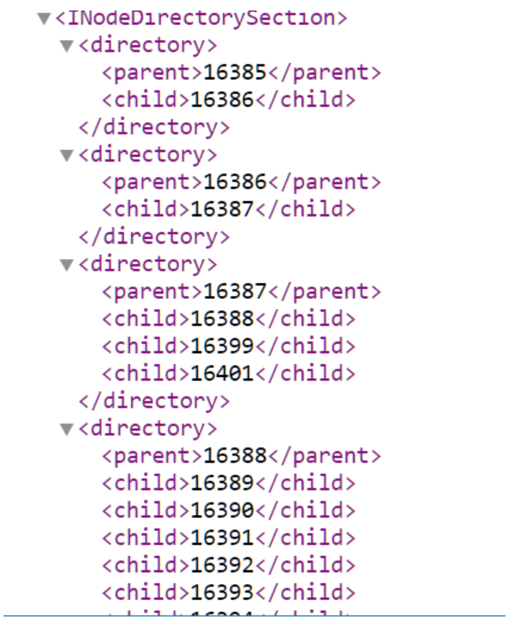
```
create database dsci551;

create user dsci551@localhost identified by "Dsci-551";

grant all privileges on dsci551.* to dsci551@localhost;
```



```
▼<inode>
   <id>16400</id>
   <type>FILE</type>
   <name>hello.txt</name>
   <replication>1</replication>
   <mtime>1675118034627</mtime>
   <atime>1675118033872</atime>
   <preferredBlockSize>134217728</preferredBlockSize>
   <permission>ubuntu:supergroup:0644</permission>
  ▼<blocks>
    ▼<block>
       <id>1073741835</id>
       <genstamp>1011</genstamp>
       <numBytes>47</numBytes>
     </block>
   </blocks>
   <storagePolicyId>0</storagePolicyId>
 </inode>
```



```
▼<INodeDirectorySection>
  ▼<directory>
     <parent>16385</parent>
     <child>16386</child>
   </directory>
  ▼<directory>
     <parent>16386</parent>
     <child>16387</child>
   </directory>
  ▼<directory>
     <parent>16387</parent>
     <child>16388</child>
     <child>16399</child>
     <child>16401</child>
   </directory>
  ▼<directory>
     <parent>16388</parent>
     <child>16389</child>
     <child>16390</child>
     <child>16391</child>
     <child>16392</child>
     <child>16393</child>
```

Tasks:

1. [25 points] Write one SQL script "create.sql" that creates the following 3 tables in a MySQL database dsci551. The script may assume that the database dsci551 already exists.

   Note that you should properly define the primary key and foreign key(if applicable) for each table and choose a suitable data type for each attribute according to the screenshot. If the above tables already exist in the database, your script should be able to recreate them.

   Your code should run without error and satisfy all requirements stated in the question.

   It's only required for you to define PK, FK and data type for each attribute. It's up to you whether to add "unique"/ "NOT NULL"/ "CHECK"/ FK CASCADE, etc.

   a. A table "inode" which stores the information about the inodes in a fsimage file. The table should have following attributes: id, type, name, replication, mtime, atime, preferredBlockSize, and permission.

   b. A table "blocks" which stores the block information for a file. The table should have the following attributes: id, inumber, genstamp, numBytes, where inumber is inode id of the file and id is a block id. Note that it is possible that a file has multiple blocks. You can assume that every block has a unique id. No need to store storagePolicyID.

   c. A table "directory" which has two attributes: parent and child, where parent is the inumber of parent directory and child is the inumber of file/directory stored under the parent directory.

2. [35 points] Write a Python script "load.py" that takes a fsimage file and stores its inode and inode directory information in the tables (in the database dsci551) you created in Task 1.

Execution format:

```
python3 load.py <fsimage.xml>
```

where <fsimage.xml> is a file system image file.

Note that your load.py should access the dsci551 database (with user = dsci551 and password = Dsci-551, as mentioned earlier) in your MySQL **on your localhost**. You can test your code by first uploading it to your EC2 instance and running it there.

Note that you are provided with a sample fsimage file but your code may be tested on additional files. This means you should not hardcode the provided fsimage file in your code. If there is no value for a particular attribute, for an integer consider it to be 0, and for string consider it to be ‘ ’. For example, in the sample fsimage92.xml inode with id=16386 does not have atime, so assign 0 to it.

You should use lxml and its xpath function to extract data from the xml file. You may use sqlalchemy, pymysql, and pandas to work with MySQL inside Python.

3. [40 points] For each of the following questions, write a MySQL query that uses the tables you created in Question 1 to answer the question. Please use queries covered during lectures.
   a. Find names of **files** which have the latest modification time. Note that it is possible that multiple files have the same modification time, please return all of them.

      format of output:
      ```
      +---------+
      | name |
      +---------+
      ```

   b. Find names of files or directories whose permissions end with 6xx. In other words, the third to the last character of the permission is 6.

      format of output:
      ```
      +---------+
      | name |
      +---------+
      ```

   c. For each file, output its id, name, and total size (in bytes).

      format of output:
      ```
      +-----------+-------------- +----------------------+
       |   id   |     name    |    total_size     |
      +-----------+-------------- +----------------------+
      ```

   d. For each directory, output its id, name, and the number of objects (files or directories) in the directory.

      format of output:
      ```
      +-----------+-------------- +---------------------+
       |   id   |     name    |    num_obj     |
      +-----------+-------------- +---------------------+
      ```

e. Find names of directories which contain at least two files.

format of output:

```
+---------+
| name |
+---------+
```

Allowed Libraries for q2: lxml, sys, sqlalchemy, pymysql, and pandas.

Submission (3 files in total):

1. For question 1: one .sql file
2. For question 2: one .py file
3. For question 3: one pdf file including all 5 queries

Notes:

1. Please strictly follow the wording of all attributes name and tables name
2. Do not zip your files when submitting your work
3. Fail to follow any submission requirements will lead to deduction of marks, please follow the instructions carefully