**CSE 5334 – Data Mining**

**Project 1**

**Design and Implementation:**

Data Structures Used : List, Dictionary, Tuple

Dictionary is used to store the values as Key and Value Pair.

1. For the Users.tsv file, we read the values which we require to construct the Data warehouse model and store the values as Key,Value Pairs

Dictionary with UserID as Key and StateID,Country as the Value

[UserID] -> (StateID,Country) Value as stored as Tuple

1. For the jobs.tsv file , we the read the values and stored in the following way

Dictionary with JobID as Key and Title as Value

[JobID] -> (Title) Value as stored as Tuple

1. For the apps.tsv file, we the columns JobID and make it as the Key and when reading the column UserID and by using these values we append the StateID , Country and Title from the first and Second dictionary .By appending the values we have constructed the Base Cuboid.

Base Cuboid is created [JobID] -> [UserID,StateID,Country,Title]

Value is stored as List

1. Slicing the Base Cuboid to get the StateID,JobID and the measure value Number of Applicants can be calculated from the this dictionary

[(StateID),JobID] -> [UserID]

Value is stored as List

1. Sorting and Ranking the items in the Sliced Dictionary to get the Top 5 Count of StateID and JobID.
2. Slicing the Base Cuboid and performing the OLAP operations to get Country,Title and the measure can be calculated from this dictionary

[(Country,Title)] -> [UserID] Values are stored in form of List

1. Sorting and Ranking the items in the Sliced Dictionary to get the Top 5 Count of TitleID and number of Applicants.

**Steps to execute the program:**

1. Unzip the folder to find the python file (datacube.py)
2. Use the command python *filename.py location of* ***users.tsv*** *location of* ***jobs.tsv*** *location of* ***apps.tsv***

(Please maintain the same order as mentioned for the input files)

1. Output for the first task will be printed with heading [Count, Location, UserID] and the output of the second task with heading [Count, TitleID]

**Snapshots:**



