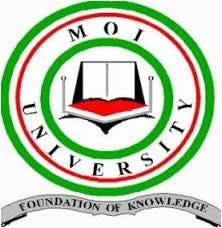
**Moi University Main Campus**

****

**School of Information Sciences**

**Department of Information Technology**

**INF 444E: DATAMINING AND ADVANCED DATABASES**

**GROUP 5 INFORMATICS**

MICHAEL GACHUJI IF/17/14

NELSON MANDELA IF/30/14

VIOLAH CHEBET IF/42/14

LAWRENCE MUTUKU IF/15/14

LEWIS RUTO IF/47/13

JUDITH NYABUTO IF/34/14

BETHSHEBA CHEPKIRUI IF/1003/14

JAMES MBITHI IF/35/14

JUDE KAWIRA IF/ 37/14

***4.10 Suppose a company wants to design a data ware house to facilitate the analysis of moving vehicles in an online analytical processing manner. The company registers huge amounts of auto movement data in the format of (Auto ID, location, speed, time). Each Auto ID represents a vehicle associated with information (e.g., vehicle category, driver category), and each location may be associated with a street in a city. Assume that a street map is available for the city.***

***(a) Design such a data warehouse to facilitate effective online analytical processing in multidimensional space.***

***(b) The movement data may contain noise. Discuss how you would develop a method to automatically discover data records that were likely erroneously registered in the data repository.***

***(c) The movement data may be sparse. Discuss how you would develop a method that constructs a reliable data warehouse despite the sparsity of data.***

***(d) If you want to drive from A to B starting at a particular time, discuss how a system may use the data in this warehouse to work out a fast route.***

**Answers:**

1. OLAP in a multidimensional space:

Driver ID

Registration

Driver category

Vehicle ID

Vehicle category

Vehicle specification

Insurance validity

Location ID

location

City

Street

Country

Auto ID

Location ID

Vehicle ID

Driver ID

Speed

Time

(b)As data moves, noise may arise. This noise can be detected through the use of appropriate data cleaning methods. Due to the constant data movement, some data sets may contain erroneous pieces of data; which causes noise. Such pieces of data are common in early stages prior to data processing. In order to automatically detect the noise in data, there are specific predefined algorithms that help in the identification of the locations of the noisy elements. A query is passed to the tuples to eliminate the noise. For instance, assuming that there is noise in the location ID, in the dimension table, the problem is identified, located and then removed before the data proceeds further.

(c) In the field of data mining, sparsity refers to thinly scattered or distributed pieces of data. In a bid to solve the issue of sparsity of data, data integration should be done. Once all the relevant data that is required to create the warehouse is obtained, the data processing and transformation of data is done. After this, the pieces of data are refreshed to keep the data warehouse up to data to perform various OLAP.

(d) According to the information provided the street map and the auto id information about driver category, the vehicle Id and the location of starting along with the time in the database. From the information already given, it is easier to find the shortest path from A to B. All the pieces of data from various starting time can be gathered to ensure a path without traffic to avoid time wastage. The final shortest path can be established by comparing the route taken starting from the source (point A) by various auto mobile to reach B(destination). On the other hand, the fastest path can be predicted by the use of the right shortest path algorithm. (Dijkstra’s algorithm). For the given data set, the location id of the source and destination, vehicle category, the speed of travel and time started is looked at. Based on comparison with other routes using the shortest path algorithm the suitable path is found from A to B.

1. Maximum number of cells possible in the **C** data cube with aggregated cells requires adding additional value to base cuboid. Therefore, the number of cells is .
2. Minimum number of cells possible in the **C** data cube with **p** distinct value except the apex value which contains a single value is

***References:1. Textbook- Data mining concepts and techniques, 3 rd. edition by Jiawei Han, Micheline Kamber, Jian Pei***

**8.** SELECT unnest(xpath(‘.//title/text();data::xml))::text AS XMLDATA FROM assign.

9. SELECT xmlelement(movie,xmlforest(filmid as ‘filmid’),

xmlforest(title as ‘title’)

xmlforest(year as ‘year’)

xmlforest(diector as ‘director’)

xmlforest(moviestars as ‘moviestars’)

)FROM movie