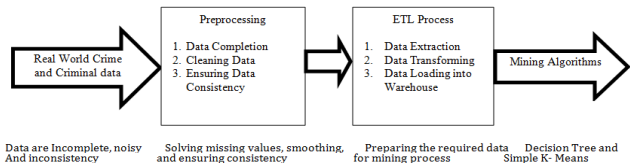






5. Data discretization: part of data reduction, replacing numerical attributes with nominal ones.
- A.** Different preprocessing techniques were used to get clean data, these include:
1. Removing outliers, some of the data is the crime and criminal datasets represent outliers and cannot be included in the analysis algorithms and techniques, so these data records were deleted from the set.
  2. Filling missing data, some criminal ages, jobs, and income were not mentioned in the tables, average and most commonly used values were used to substitute these missing values.
  3. Data reduction using normalization and aggregation.

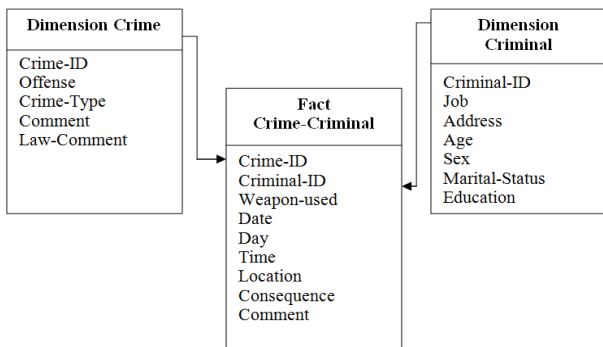
**B.** The process is shown in figure 2.



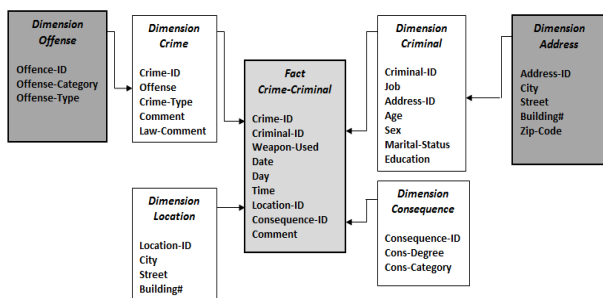
**Figure 2: Crime and Criminal Data Preprocessing**

#### 4. Proposed Frame Work Design

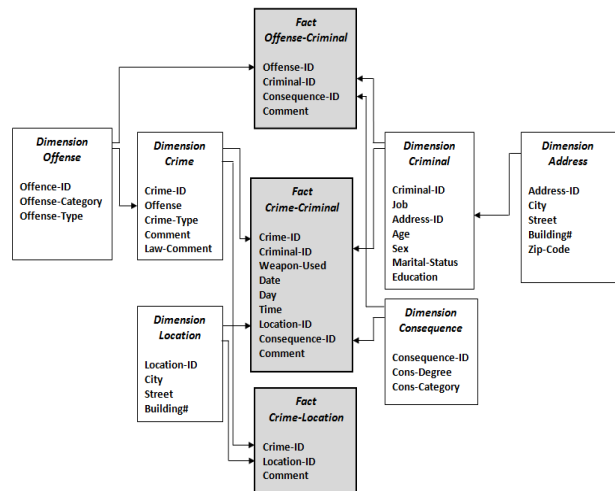
Three different design models are available for data warehouse, they are Star, snow flake and galaxy model[1]. Each model has its advantages and drawbacks. It is clear that scanning the entire table of n records requires  $O(n)$  time complexity, whereas scanning the fact table Crime\_criminal shown in figure3 requires a time complexity highly dependent on the number of records in both dimensions Crime and Criminal. Hence using normalized schema as shown in figure 4 and many fact tables as shown in figure 5 will have a great effect on the time complexity required to carry out the analysis process.



**Figure 3: DW Star Model for the Proposed System**



**Figure 4: DW Snowflake (normalized) Model for the Proposed System**



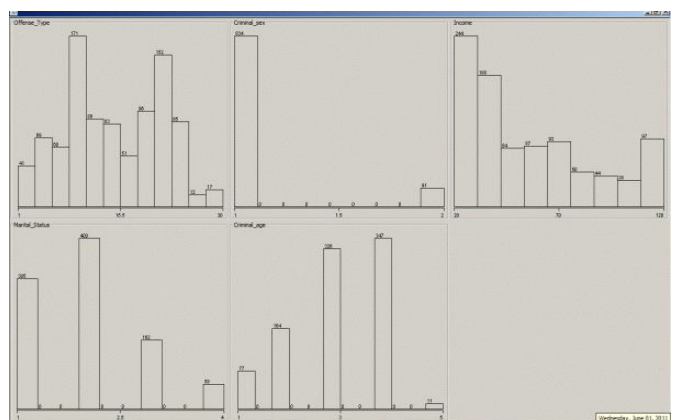
**Figure 5: DW Galaxy Model for the Proposed System**

**Table 1:** shows sample of the crime and criminal data.

**Table 1: Crime and Criminal Sample Data**

Crime ID	Offense Type	Day of the Week: 1=Sun*	Criminal ID	Criminal Sex 1=Male	Criminal Income In K	Criminal Marital Status 1=Single	Criminal Age Category 1=Age<20
24	1	6	271	1	120	4	4
36	1	5	50	1	20	2	1
123	1	5	408	1	50	2	1
218	1	6	586	1	35	2	2
231	1	7	10	1	80	1	3
242	1	2	286	1	20	2	3
316	1	5	554	1	65	2	2
364	1	7	165	1	20	1	4
404	1	6	575	1	40	1	3
444	1	6	446	1	40	2	3
592	1	6	356	1	40	1	3
602	1	7	74	1	40	1	4
678	1	7	411	1	55	2	2
686	1	7	315	1	40	2	3
945	1	6	31	1	30	3	4
955	1	6	194	2	50	1	2
102	2	1	165	1	20	1	4
106	2	1	510	1	50	2	1
116	2	2	255	1	35	3	3
211	2	6	468	1	40	2	3
212	2	6	420	1	110	1	4
415	2	1	194	2	50	1	2
596	2	6	45	1	20	2	2
762	2	6	481	1	40	1	4
773	2	6	310	2	20	2	2

Applying the preprocessing algorithms and techniques on the collected data in the dimensions and fact tables shown in figures 3,4 and 5 gave the distributions and histograms in figure6, were WEKA software was used to get such distributions.



**Figure 6: Distribution of Crime and Criminal data with Different Attributes**

## 5. Results and Analysis

Three different DW models including Star, Snow Flake and Galaxy were used to design the required repository for the logged data, this will help in improving the analysis performance and help ensuring data privacy. Different mining techniques were used to analyze the logged data, these include: Clustering, Association and classification with different algorithms.

### 5.1. Clustering

K Means clustering algorithm was used to group criminal objects as shown in Table 2.

**Table 2: Clustering Technique Results**

Attribute	Full Data 925 records	Cluster #	
		0 91 records	1 834 records
sex	1.0984	2	1
Income	55.78	54.01	55.97
Marital Status	1.9708	2.0659	1.9604
Age	3.0551	2.8791	3.0743

#### Clustered Instances

0 91 ( 10%)  
 1 834 ( 90%)

### 5.2. Classification

#### REPTree algorithm

```

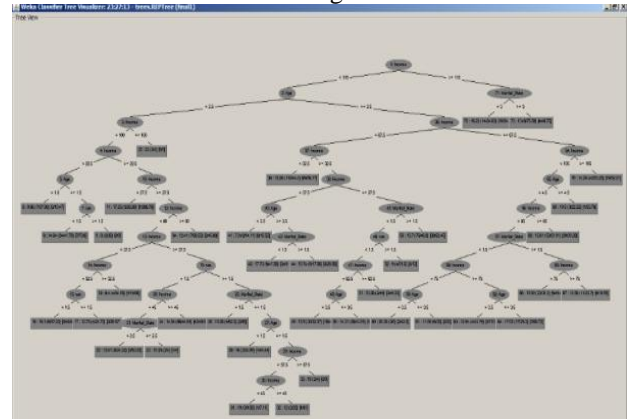
Income < 115
| Age < 2.5
| | Income < 100
| | | Income < 22.5
| | | | Age < 1.5 : 9.86 (7/37.96) [3/10.47]
| | | | Age >= 1.5
| | | | | sex < 1.5 : 14.64 (24/41.79) [7/73.6]
| | | | | sex >= 1.5 : 8 (2/36) [0/0]
| | | | Income >= 22.5
| | | | | Income < 27.5 : 17.25 (3/26.89) [1/266.78]
| | | | | Income >= 27.5
| | | | | | Income < 80
| | | | | | | Income < 37.5
| | | | | | | | Income < 32.5
| | | | | | | | | sex < 1.5 : 18.1 (6/97.25) [5/48.65]
| | | | | | | | | sex >= 1.5 : 12.75 (4/24.75) [3/50.92]
| | | | | | | | | Income >= 32.5 : 8.4 (4/54.19) [1/18.06]
| | | | | | | | Income >= 37.5
| | | | | | | | | sex < 1.5
| | | | | | | | | | Income < 45
| | | | | | | | | | | Marital_State < 2.5 : 13.61 (6/24.22) [3/63.33]
| | | | | | | | | | | | Marital_State >= 2.5 : 10.33 (2/4) [1/4]
| | | | | | | | | | | | | Income >= 45 : 14.54 (86/44.61) [42/48.01]
| | | | | | | | | | | | | sex >= 1.5
| | | | | | | | | | | | | | Marital_State < 1.5 : 10.28 (4/62.5) [2/80]
| | | | | | | | | | | | | | Marital_State >= 1.5
| | | | | | | | | | | | | | | Age < 1.5 : 16 (3/24.89) [1/44.44]
| | | | | | | | | | | | | | | Age >= 1.5
| | | | | | | | | | | | | | | | Income < 57.5
    
```

```

Income < 45 : 19 (3/9.56) [1/7.11]
Income >= 45 : 12 (2/25) [1/81]
Income >= 57.5 : 19 (2/4) [0/0]
Income >= 80 : 12.41 (7/58.53) [2/45.68]
Income >= 100 : 23 (2/4) [0/0]
Age >= 2.5
| Income < 67.5
| | Income < 32.5 : 15.26 (110/44.4) [69/56.17]
| | Income >= 32.5
| | | Income < 37.5
| | | | Age < 3.5 : 7.9 (8/14.11) [3/15.52]
| | | | Age >= 3.5
| | | | | Marital_State < 1.5 : 17.73 (6/41.56) [3/49.11]
| | | | | Marital_State >= 1.5 : 10.74 (6/17.56) [3/28.56]
| | | | | Income >= 37.5
| | | | | | Marital_State < 1.5
| | | | | | | sex < 1.5
| | | | | | | | Income < 62.5
| | | | | | | | | Age < 3.5 : 15.9 (30/52.57) [16/44.5]
| | | | | | | | | Age >= 3.5 : 14.31 (26/43.91) [14/58.92]
| | | | | | | | | Income >= 62.5 : 13.56 (2/49) [3/40.33]
| | | | | | | | | sex >= 1.5 : 9 (4/21.5) [2/13]
| | | | | | | | | | Marital_State >= 1.5 : 13.7 (75/46.6) [38/52.45]
| | | | | | | | | | Income >= 67.5
| | | | | | | | | | | Income < 105
| | | | | | | | | | | Age < 4.5
| | | | | | | | | | | | Income < 85
| | | | | | | | | | | | | Marital_State < 1.5
| | | | | | | | | | | | | | Income < 75
| | | | | | | | | | | | | | | Age < 3.5 : 20.38 (3/0) [2/42.5]
| | | | | | | | | | | | | | | Age >= 3.5 : 11.06 (6/30) [2/52]
| | | | | | | | | | | | | | | Income >= 75
| | | | | | | | | | | | | | | | Age < 3.5 : 13.94 (4/43.19) [3/110.56]
| | | | | | | | | | | | | | | | Age >= 3.5 : 17.05 (17/29.3) [7/66.72]
| | | | | | | | | | | | | | | | | Marital_State >= 1.5
| | | | | | | | | | | | | | | | | | Income < 75 : 15.92 (23/31.2) [9/49.41]
| | | | | | | | | | | | | | | | | | Income >= 75 : 13.56 (11/23.7) [8/19.59]
| | | | | | | | | | | | | | | | | | Income >= 85 : 15.81 (53/51.91) [20/39.28]
| | | | | | | | | | | | | | | | | | Age >= 4.5 : 10.5 (3/22.22) [1/53.78]
| | | | | | | | | | | | | | | | | | Income >= 105 : 14.29 (42/53.25) [19/52.21]
Income >= 115
| Marital_State < 3 : 16.22 (14/24.63) [10/54.2]
| Marital_State >= 3 : 13 (6/75.58) [4/48.75]
    
```

Size of the tree : 73

The decision tree is show in figure 5.



**Figure 5: Decision Tree for Classification Process**

## 6. Conclusion

Good mining results can be achieved when the historical data are big enough, and in crime analysis this is highly true. Anyhow, samples of about one thousand crime records and more than six hundred for criminals are enough to get a good result in the proposed model. WEKA (Waikato Environment for Knowledge Analysis) and Excel software were used to analyze the collected crime and the criminal data.

First of all, the collected data were preprocessed to fill in the missing attributes and remove outliers and then data were normalized and transformed into formats suitable for analysis purposes. Table 1 shows sample of the data after preprocessing. It is clear that data in table 1 can be very well fitted for analysis using decision tree and clustering algorithms. Three DW models were used as repositories for the data highly affected the analysis process algorithms performance since the whole schema is normalized and data reduction technique is applied when using Galaxy model as data repository, i.e. time complexity is highly improved.

The results from clustering algorithm showed that criminals can be divided into two groups or clusters each has its own attribute values for age, gender, marital status and job and from this result we can predict any other unknown object of type criminal. Rules and decision tree given in figure 5 are very well suitable for criminal classification. From the Decision Tree given in figure 5 it is clear that attributes income and marital status have the higher priority affecting the classification process from which we can conclude that these two attributes can be used at the top level of the decision tree to classify criminal into groups. The paths for different types of offenses depending on different crime and criminal attributes, this will help in identifying what attributes highly affect a specific type of offense. Entropy and information gain locate the attributes highly affecting the results at the top of the tree.

## References

- [1] Jiawei Han and Micheline Kamber "Data Mining: Concepts and Techniques" 2<sup>nd</sup> ed., Morgan Kaufmann, 2006.
- [2] M. Steinbach, P.-N. Tan and V. Kumar, Introduction to Data Mining, Addison-Wesley, 2006. ISBN: 0-321-32136-7
- [3] M. H. Dunham, Data Mining: Introductory and Advanced Topics, Prentice Hall, 2002.
- [4] D. J. Hand, H. Mannila, and P. Smyth, Principles of Data Mining, MIT Press, 2001.
- [5] Deborah Osborne, MA, Susan Wernicke, MS, "Introduction to Crime Analysis: Basic Resources for Criminal Justice Practice, The Haworth Press, New York, London, Oxford, 2003.
- [6] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2<sup>nd</sup> ed., 2005, ISBN 0-12-088407-0
- [7] Haider k. and Kadhim Aljanabi, "Crime Data Analysis Using Data Mining Techniques To Improve Crimes Prevention Procedures", ICIT2010, October 2010, University of Kufa, Iraq.

- [8] Austin Police Department Office, <http://www.ci.austin.tx.us/police/crime.htm>
- [9] Derek J. Paulsen, Sean Bair, and Dan Helms Tactical Crime Analysis: Research and Investigation, 2009.
- [10] Hsinchun Chen, Homa Atabakhsh, Tim Petersen, "Visualization for Crime Analysis", Proceedings of The National Conference on Digital Government Research CiteSeerx, COPLINK, 2006.
- [11] Tianyi Wu, Yuguo Chen and Jiawei Han, "Association Mining in Large Databases: A Re-Examination of Its Measures", in Proc. 2007 Int. Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD'07), Warsaw, Poland, Sept. 2007.

