ISSN 2278-3091
Volume 9, No.5, September - October 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse313952020.pdf
https://doi.org/10.30534/ijatcse/2020/313952020

An Optimal Warehouse Design for Crime Dataset

Haider Alsharqi[1], Kadhim B. S. Aljanabi[2]
[1]Information Technology Research and Development Centre, University of Kufa, Najaf, Iraq
haider.alsharqi@uokufa.edu.iq
[2]Department of Computer Science, Faculty of Computer Science and Mathematics, University of Kufa, Najaf, Iraq
kadhim.aljanabi@uokufa.edu.iq

ABSTRACT
Data science and analytics represent one of the most emerging fields nowadays. Collecting, storing and analyzing the data are challenging issues in the field since they require the most advanced techniques and technologies. Data Warehouse and Data Marts represent some solutions for collecting, storing and accessing the data. Good Warehouse design leads to better analysis results.

Among different application fields of the data, crime data is an important and complex discipline that contains a number of complex relationships between its contents, a wide range of applications and its crucial importance.

The aim of the work in this paper is building an optimal Data warehouse for crime dataset using real crime data collected from the internet. Among the different DW modules available in this field galaxy module is used in this work. The data warehouse will support the decision-making process for lawmaker and police departments by understanding crime subjects, and statistics that allow them to track actions, foretell the probability of occurring crimes and efficiently use supplies which are inverted in this paper.

The proposed design of the DW shows more reliability, better storing and accessing capabilities and lower anomalies among the other designs. The proposed design was supported with a crime database design to remove heterogonous of the data and to apply some preprocessing issues from which they require data is extracted, transformed and loaded (ETL) into the warehouse.

Finally, more than six million high quality, clean, and preprocessed of crime records data are available for the researchers.

Key words: Data warehouse, Database, Data Reduction, Preprocessing.

1. INTRODUCTION

The fast growth of cloud computing and data acquisition and storage technologies, from business and research centres of governments and different organizations, has led to a vast number of unprecedented, complex from data that has been gathered and produced publicly available [1]. It has become more critical to extort meaningful information and provide new insights for understanding patterns from such data stores. Data mining can efficiently address the difficulties of data that are too enormous, unstructured, and fast-moving that not handle by traditional approaches [2]. Data mining is an innovative, interdisciplinary, and growing research field, which can build models and techniques across several areas for inferring useful information and hidden patterns from data [3],[4].The data mining techniques, like clustering, classification and association, can be used for data analysis and prediction by beneficial extract information from raw data.

1- Clustering: is unsupervised learning of data mining technique and is an automatic manner in which the data divided into groups whose segments are similar to each other these are called clusters. For similarity between elements in each cluster, different measures can reflect, such as the distance measure, which uses in the K-Means algorithm, this applies for clustering and samples that are closest to each other can be held as one cluster. So, in image and video databases, clustering can use to discover exciting patterns also characteristics and support content-based retrievals of images and videos using low-level features such as colour, shape, histogram descriptions, texture [5]. There are several algorithms for clustering can divide into two groups:

A. Traditional Clustering Algorithms: algorithms based on one of the features like Partition, Hierarchy, Fuzzy Theory, Distribution, Density, Graph Theory, Grid, Fractal Theory and Model.

B. Modern Clustering Algorithms: algorithms based on one of the features like Kernel, Ensemble, Swarm Intelligence, Quantum Theory, Spectral Graph Theory, Affinity Propagation and Density and Distance. Also, the algorithm for Clustering Spatial Data, Clustering Algorithm for Data Streams and Clustering Algorithm for Large-Scale Data [6].

2- Classification: is very strictly associated with the clustering and assigns to as supervised learning. A classification is an approach in data mining which the algorithm learning from the input data and then applies this information to classify new results. The classification has been examined widely by the database and Artificial Intelligence areas. Some well-known methods of classification are Decision Tree, k-Nearest Neighbor, DNF Rules, Neural Networks and Bayesian Classifiers [7],[8].

3- Association: is a process of finding relationships between different attributes in large data sets in various types of databases. These attributes maybe 0 or 1, or they may be quantitative. The concept in association rule is finding the kind of causalities between the values of the various attributes. Association rule mining includes the use of machine learning models to analyze data for patterns, or co-occurrence, in a database. It identifies many if-then relationships, which are called association rules. Association rule generation has considerable importance in data mining because of the ability of its use as an essential mechanism for knowledge finding. For example, a supermarket which the data that records for the different events is the sets of items bought by each customer. For that, it may be helpful to find how the shopping behaviour of one element influences the shopping behaviour of another, Association Rules support for detecting such relationships correctly. Before-mentioned data may use to make target marketing choices. It can also be generalized to classify on big data [9].

There are various types of mining algorithms using. Despite such diversity, some methods are more frequently use [10]. In addition to these methods, a data warehouse, invented by Bill Inmon, is a procedure used in data analysis, database format to analyzing and to report which recognize as OLAP (Online Analytical Processing). From this process, data analysis and doing statistics can be quickly and accurately with refreshed data. Data warehouse structure is a database form that focused on the utilization of statistics. With a data warehouse, architecture reporting process becomes faster, interactive and also can be in real-time [11]. Data warehouse evolved from a normalized database by using ETL (Extraction Transformation and Loading). So, the applications of data mining used to extract knowledge from data like medicine, marketing, weather, crime and various complex data.

There are many models in data warehouse design which represented star, snowflake and galaxy schema. These schemas are created based on reporting needed. in the following sentences, we list the standard schemas with a brief description for each schema:

1-Star schema: is the most straightforward and widely used form of data warehouse schemas, include one or more fact tables referencing any dimension tables. Star schema is more efficient with working on simpler queries,

2-Snowflake schema: is similar to the star schema. However, in snowflake dimensions are arrangement in various related tables with multiple levels of table relations,

3-Galaxy schema: is more complicated of star and snowflake schemas, it contains multiple fact tables connected with various dimension tables that sorted in one level or more, seem like groups of stars [12].

Due to continuous urbanization and growing populations, especially in enormous societies, increasing Violent crimes and accidents recently led to evidence of crucial and different databases that resulted in needing for getting valuable information to analyze crimes by using diverse technologies [13]. Data mining methods are applied to crime data to identify the complexity of the relationship between the criminals and crime pattern by using clustering and classification algorithms with a suitable data structure.

2. RELATED WORKS

The academic literature on data warehouse design has revealed the emergence of several contrasting themes.

Agapito et al. (2020) [14] design COVID-WAREHOUSE where they integrate and save the COVID-19 data, several pollutions and weather data made available in Italy. Also, the data warehouse supports Public Health to understand how the pandemic is spreading in time and geographic area and to associate the pandemic to pollution and climate data in a particular region.

ANUSHA and Jyothi (2020) [15] design a data warehouse for a medical information system to examine the process of data, take decisions, foresee diseases and find cures with the help of data warehouse.

Quitaleg and Ortiz (2020) [16] design and Development of Data Warehouse of Highland Vegetable Crops, this warehouse offers an efficient method for analysis and statistics to the big data in agriculture. It takes data about vegetables, farmers, consumers and other factors and requirements concerning agricultural production. A data warehouse introduces as a solution to agricultural data issues.

A.Abdo et al. (2019) [17] design a system for crime prediction by collecting data from Egyptian forensic medicine and create a data warehouse for this data to apply data mining techniques on it, Where the system achieved acceptable results about 98%.

Ari Setyawan et al. (2019) [11] design a data warehouse for an insurance company will help to perform data analysis and to report better and more efficient.it is helps for analyzing the results of insurance sales and used as a reference to the administration to make decisions.

Prabawa et al. (2019) [18] building data warehouse for e-travel company by using the fourth steps dimensional modelling methodology, the design using snowflake scheme become the solution to predict business trends, maintain quality, enhance competitiveness, and exist in the long run for companies.

Sudarmojo et al. (2018) [19] design a data warehouse for a library to analyze data about the transaction process for getting smart decisions in similar service feature evaluation in libraries.

Farooqui et al. (2018) [20] introduce a methodology for the construction of a data warehouse for a medical information system; this data warehouse will help to improve the data analysis and assisting clinical managers to identify decisive patterns, diseases, and their support by enhancing the decision making.

Choo and Chua (2018) [21] implement and design a data warehouse for the semi-structured research literature mining, this study shows data warehouse able to support for exposing hidden pattern or trend of certain aspects of the research literature data such as the keyword locus of a journal. Also discover information about the authors who had cooperated before in writing research literature.

Sutedja et al. (2018) [22] design data warehouse to support active student management by using four stages used by Ralph Kimball in designing a data warehouse. This data warehouse will help the university to analyze active student and make decisions in the student area.

3. METHODOLOGY

The proposed approach to design and implement a data warehouse for crime dataset contains the following six stages:

1. Data collection
2. Convert CSV file to SQL table
3. Data Preprocessing
4. Database Design (an intermediate stage to overcome data heterogeneous)
5. Data warehouse design and the second stage of preprocessing
6. Evaluation and analysis

A Proposed Warehouse Design Phases are shown in Figure (1).

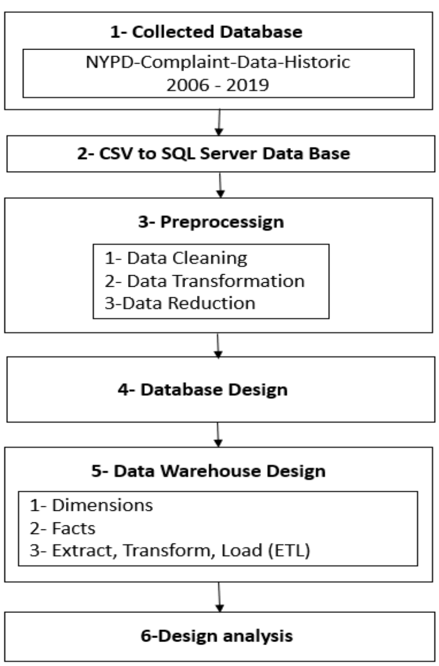


Figure 1: Proposed Warehouse Design Phases

3.1 Collected Data

The Crime dataset used in this paper is real data on New York City (NYPD Complaint Data Historic) [23]. Which contain 35 columns with 6,500,871 rows. In CSV file, the data describe the type of crime "(OFNS_DESC, PD_DESC, CRM_ATPT_CPTD_CD, LAW_CAT_CD)", suspect "(SUSP_AGE_GROUP,SUSP_RACE,SUSP_SEX)", victims "(VIC_AGE_GROUP,VIC_RACE,VIC_SEX)" and the location "(BORO_NM, Latitude, Longitude)" for each crime reported on it in NYPD from 2006 to the end of 2019. Sample of collected raw data is shown in Table (1), Table (2) shows the detailed description of the data attributes.

Table 1: Sample of Collected Raw Data from CSV File

	CMPLNT_NUM	CMPLNT_FR_DT	CMPLNT_FR_TM	CMPLNT_TO_DT	CMPLNT_TO_TM	ADDR_PCT_CD	RPT_DT	KY_CD	OFNS_DESC	PD_CD	PD_DESC	CRM_ATPT_CPTD_CD	LAW_CAT_CD	BORO_NM	LOC_OF_OCCUR_DESC	PREM_TYP_DESC	JURIS_DESC
2	522575447	8/29/2006	13:00:00			43	8/30/2006	578	HARRASSMENT 2	638	HARASSMENT,SUBD 3,4,5	COMPLETED	VIOLATION	BRONX	INSIDE	RESIDENCE - APT. HOUSE	N.Y. POLICE DEPT
3	403507361	11/5/2006	11:00:00	11/5/2006	17:40:00	66	11/5/2006	107	BURGLARY	221	BURGLARY, RESIDENCE, DAY	COMPLETED	FELONY	BROOKLYN	INSIDE	RESIDENCE - APT. HOUSE	N.Y. POLICE DEPT
4	631420068	9/8/2006	23:30:00	9/9/2006	0:01:00	106	9/9/2006	347	INTOXICATED & IMPAIRED DRIVING	905	INTOXICATED DRIVING,ALCOHOL	COMPLETED	MISDEMEANOR	QUEENS	FRONT OF	STREET	N.Y. POLICE DEPT
5	995609899	12/13/2011	18:40:00	12/13/2011	18:49:00	79	12/13/2011	341	PETIT LARCENY	333	LARCENY,PETIT FROM STORE-SHOPL	COMPLETED	MISDEMEANOR	BROOKLYN	INSIDE	CHAIN STORE	N.Y. POLICE DEPT
6	480667624	8/14/2009	4:20:00			30	8/14/2009	113	FORGERY	729	FORGERY,ETC.,UNCLASSIFIED-FELO	COMPLETED	FELONY	MANHATTAN		STREET	N.Y. POLICE DEPT

	JURISDICTION_CODE	PARKS_NM	HADEVELOPT	HOUSING_PSA	X_COORD_CD	Y_COORD_CD	SUSP_AGE_GROUP	SUSP_RACE	SUSP_SEX	TRANSIT_DISTRICT	Latitude	Longitude	Lat_Lon	PATROL_BORO	STATION_NAME	VIC_AGE_GROUP	VIC_RACE	VIC_SEX
2	0	NA		NA	1018029	240747		UNKNOWN	M		40.827414	-73.877946	4051, -73.87	PATROL BORO BRONX		25-44	BLACK HISPANIC	F
3	0	NA		NA	982556	171385					40.637097	-74.006105	6864, -74.00	PATROL BORO BKLYN SOUTH		45-64	ASIAN / PACIFIC ISLANDER	F
4	0	NA		NA	1028213	186786					40.67926	-73.8415	0229, -73.84	PATROL BORO QUEENS SOUTH			UNKNOWN	E
5	0	NA		NA	1000788	189718					40.687402	-73.940369	1619, -73.94	PATROL BORO BKLYN NORTH			UNKNOWN	D
6	0	NA		NA	1000029	242245					40.831576	-73.942983	6128, -73.94	PATROL BORO MAN NORTH			UNKNOWN	E

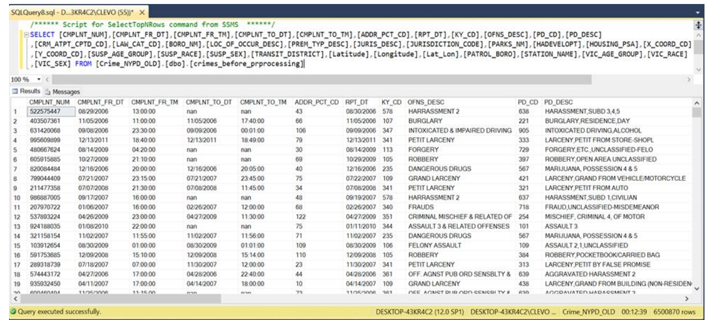
Table 2: Attribute Description of The Collected Raw Data

	Column Name	Column Description
1	CMPLNT_NUM	Randomly generated persistent ID for each complaint
2	CMPLNT_FR_DT	Exact date of occurrence for the reported event (or starting date of occurrence, if CMPLNT_TO_DT exists)
3	CMPLNT_FR_TM	Exact time of occurrence for the reported event (or starting time of occurrence, if CMPLNT_TO_TM exists)
4	CMPLNT_TO_DT	Ending date of occurrence for the reported event, if exact time of occurrence is unknown
5	CMPLNT_TO_TM	Ending time of occurrence for the reported event, if exact time of occurrence is unknown
6	ADDR_PCT_CD	The precinct in which the incident occurred
7	RPT_DT	Date event was reported to police
8	KY_CD	Three digit offense classification code
9	OFNS_DESC	Description of offense corresponding with key code
10	PD_CD	Three digit internal classification code (more granular than Key Code)
11	PD_DESC	Description of internal classification corresponding with PD code (more granular than Offense Description)
12	CRM_ATPT_CPTD_CD	Indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely
13	LAW_CAT_CD	Level of offense: felony, misdemeanor, violation
14	BORO_NM	The name of the borough in which the incident occurred

15	LOC_OF_OCCUR_DESC	Specific location of occurrence in or around the premises; inside, opposite of, front of, rear of
16	PREM_TYP_DESC	Specific description of premises; grocery store, residence, street, etc.
17	JURIS_DESC	Description of the jurisdiction code
18	JURISDICTION_CODE	Jurisdiction responsible for incident. Either internal, like Police(0), Transit(1), and Housing(2); or external(3), like Correction, Port Authority, etc.
19	PARKS_NM	Name of NYC park, playground or greenspace of occurrence, if applicable (state parks are not included)
20	HADEVELOPT	Name of NYCHA housing development of occurrence, if applicable
21	HOUSING_PSA	Development Level Code
22	X_COORD_CD	X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
23	Y_COORD_CD	Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
24	SUSP_AGE_GROUP	Suspect's Age Group
25	SUSP_RACE	Suspect's Race Description
26	SUSP_SEX	Suspect's Sex Description
27	TRANSIT_DISTRICT	Transit district in which the offense occurred.
28	Latitude	Midblock Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
29	Longitude	Midblock Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
30	Lat_Lon	Geospatial Location Point (Latitude and Longitude combined)
31	PATROL_BORO	The name of the patrol borough in which the incident occurred
32	STATION_NAME	Transit station name
33	VIC_AGE_GROUP	Victim's Age Group
34	VIC_RACE	Victim's Race Description
35	VIC_SEX	Victim's Sex Description

3.2 Converting CSV file to SQL table

Converted the CSV file to SQL Server database by implemented python 3.6 code with (pyodbc and pandas) packages, this language was used because it is open-source language, can be used for any platform operating system and also suitable for data mining and big data [24].The SQL table is shown in Figure (2).


SQLQuery8.sql - D...3KR4C2\CLEVO (55))*

```
/****** Script for SelectTopNRows command from SSMS ******/
SELECT [CMPLNT_NUM],[CMPLNT_FR_DT],[CMPLNT_FR_TM],[CMPLNT_TO_DT],[CMPLNT_TO_TM],[ADDR_PCT_CD],[RPT_DT],[KY_CD],[OFNS_DESC],[PD_CD],[PD_DESC]
,[CRM_ATPT_CPTD_CD],[LAW_CAT_CD],[BORO_NM],[LOC_OF_OCCUR_DESC],[PREM_TYP_DESC],[JURIS_DESC],[JURISDICTION_CODE],[PARKS_NM],[HADEVELOPT],[HOUSING_PSA],[X_COORD_CD]
,[Y_COORD_CD],[SUSP_AGE_GROUP],[SUSP_RACE],[SUSP_SEX],[TRANSIT_DISTRICT],[Latitude],[Longitude],[Lat_Lon],[PATROL_BORO],[STATION_NAME],[VIC_AGE_GROUP],[VIC_RACE]
,[VIC_SEX] FROM [Crime_NYPD_OLD].[dbo].[crimes_before_prprocessing]
```

	CMPLNT_NUM	CMPLNT_FR_DT	CMPLNT_FR_TM	CMPLNT_TO_DT	CMPLNT_TO_TM	ADDR_PCT_CD	RPT_DT	KY_CD	OFNS_DESC	PD_CD	PD_DESC
1	522575447	08/29/2006	13:00:00	nan	nan	43	08/30/2006	578	HARRASSMENT 2	638	HARASSMENT,SUBD 3,4,5
2	403507361	11/05/2006	11:00:00	11/05/2006	17:40:00	66	11/05/2006	107	BURGLARY	221	BURGLARY,RESIDENCE,DAY
3	631420068	09/08/2006	23:30:00	09/09/2006	00:01:00	106	09/09/2006	347	INTOXICATED & IMPAIRED DRIVING	905	INTOXICATED DRIVING,ALCOHOL
4	995609899	12/13/2011	18:40:00	12/13/2011	18:49:00	79	12/13/2011	341	PETIT LARCENY	333	LARCENY,PETIT FROM STORE-SHOPL
5	480667624	08/14/2009	04:20:00	nan	nan	30	08/14/2009	113	FORGERY	729	FORGERY,ETC.,UNCLASSIFIED-FELO
6	605915885	10/27/2009	21:10:00	nan	nan	69	10/29/2009	105	ROBBERY	397	ROBBERY,OPEN AREA UNCLASSIFIED
7	820084484	12/16/2006	20:00:00	12/16/2006	20:05:00	40	12/16/2006	235	DANGEROUS DRUGS	567	MARIJUANA, POSSESSION 4 & 5
8	799044409	07/21/2007	23:15:00	07/21/2007	23:45:00	75	07/22/2007	109	GRAND LARCENY	421	LARCENY,GRAND FROM VEHICLE/MOTORCYCLE
9	211477358	07/07/2008	21:30:00	07/08/2008	11:45:00	34	07/08/2008	341	PETIT LARCENY	321	LARCENY,PETIT FROM AUTO
10	986887005	09/17/2007	16:00:00	nan	nan	48	09/19/2007	578	HARRASSMENT 2	637	HARASSMENT,SUBD 1,CIVILIAN
11	207970722	01/06/2007	16:00:00	02/26/2007	12:00:00	68	02/26/2007	340	FRAUDS	718	FRAUD,UNCLASSIFIED-MISDEMEANOR
12	537893224	04/26/2009	23:00:00	04/27/2009	11:30:00	122	04/27/2009	351	CRIMINAL MISCHIEF & RELATED OF	254	MISCHIEF, CRIMINAL 4, OF MOTOR
13	924188035	01/08/2010	22:00:00	nan	nan	75	01/11/2010	344	ASSAULT 3 & RELATED OFFENSES	101	ASSAULT 3
14	321158154	11/02/2007	11:55:00	11/02/2007	11:56:00	71	11/02/2007	235	DANGEROUS DRUGS	567	MARIJUANA, POSSESSION 4 & 5
15	103912654	08/30/2009	01:00:00	08/30/2009	01:01:00	109	08/30/2009	106	FELONY ASSAULT	109	ASSAULT 2,1,UNCLASSIFIED
16	591753685	12/09/2008	15:10:00	12/09/2008	15:14:00	110	12/09/2008	105	ROBBERY	384	ROBBERY,POCKETBOOK/CARRIED BAG
17	289318739	07/18/2007	07:00:00	11/30/2007	12:00:00	23	11/30/2007	341	PETIT LARCENY	313	LARCENY,PETIT BY FALSE PROMISE
18	574443172	04/27/2006	17:00:00	04/28/2006	22:40:00	44	04/28/2006	361	OFF. AGNST PUB ORD SENSBLTY &	639	AGGRAVATED HARASSMENT 2
19	935932450	04/11/2007	17:00:00	04/14/2007	18:00:00	10	04/14/2007	109	GRAND LARCENY	438	LARCENY,GRAND FROM BUILDING (NON-RESIDEN

Query executed successfully. DESKTOP-43KR4C2 (12.0 SP1) DESKTOP-43KR4C2\CLEVO ... Crime_NYPD_OLD 00:12:39 6500870 rows

Figure 2: Sample of Data from SQL Table

3.3 Data Preprocessing

Data preprocessing is a significant stage of data analysis because real-world data is impure, contain a lot of missing values or duplicating data, and high-performance mining methods expect quality data [25]. So, we will apply three steps of data preprocessing to enhancement the quality of the dataset.

3.3.1 Data Cleaning

- Deleting(CMPLNT_TO_DT,CMPLNT_TO_TM, JURIS_DESC,URISDICTION_CODE,PARKS_NM) colums because these columns not important in future analysis.In the other hand, deleted the columns that have Too many missing values(HADEVELOPT,HOUSING_PSA,TRANSIT_DISTRICT, STATION_NAME) also deleted the columns that have duplicated data(X_COORD_CD, Y_COORD_CD, Lat_Lon).
- Deleting every row with missing value in columns (CMPLNT_FR_DT, CMPLNT_FR_TM, OFNS_DESC, CRM_ATPT_CPTD_CD, BORO_NM, Latitude, Longitude, PATROL_BORO),and convert every missing value in these columns (LOC_OF_OCCUR_DESC, PREM_TYP_DESC, SUSP_AGE_GROUP, SUSP_RACE, SUSP_SEX, VIC_AGE_GROUP, VIC_RACE, VIC_SEX) to (UNKNOWN) value. And also add same (OFNS_DESC) with (PD_DESC) to rows that have missing value in (PD_DESC) column.

3.3.2 Data Transformation

In this step, we convert "(ADDR_PCT_CD, KY_CD, PD_CD, Latitude, Longitude)" columns to numerical data from string format after export it from CSV file to SQL Server Database. Also converting "(CMPLNT_FR_DT, CMPLNT_FR_TM, RPT_DT)" columns into date time format.

3.3.3 Data Reduction

Rounding the Latitude and Longitude columns into two decimals numbers to reduce the distinct values in these columns. So, it can be used with Patrol name "(P_name)" to find the Approximate location for crime.

After the preprocessing stage, the SQL table contains 23 columns and 6,435,235 rows. The column names have been changed into more meaningful names. Sample of data after preprocessing phase is shown in Figure (3).

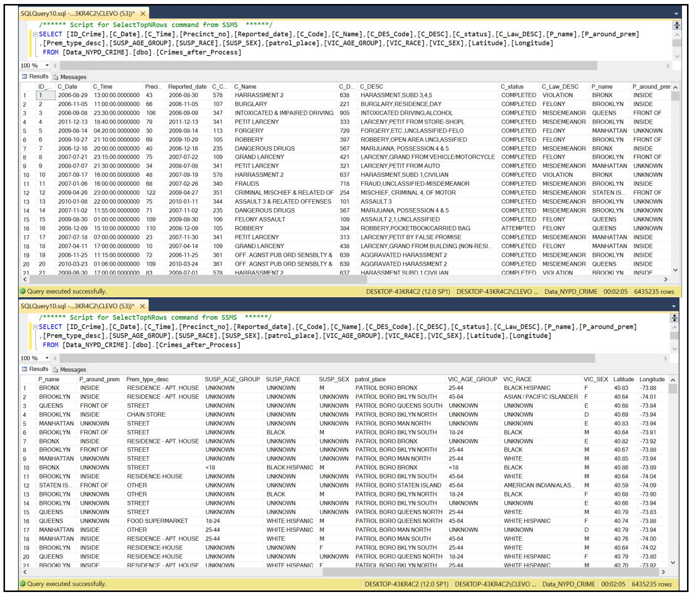


Figure 3: Sample of Data After Preprocessing Phase

3.4 Database Design

In this stage, data was converted into a query-based format suitable for data Loading from the database to DW. Three different DB schemas have been designed and created to fulfill the ETL requirements. Each ERD meets specific needs for decision-makers, and ERD is a well-known process for executing relational database [26]. An ERD contains several entities and connectors that imagine two crucial information; The main tables within the system and the inter-relationships between these tables, which are very important in designing different fact tables in the proposed warehouse. Figure (4) shows the first design for ER. This DB diagram consists of (Crimes, ListCrimesTypes, List_Places, Info_Premises, ListSuspectClasses,ListVictimClasses, Patrol) tables with relations(Crime_Places,Crime_Patrol,ListPlaces_InfoPremises,Crime_Suspect,Crime_Victim,Suspect_Victim).

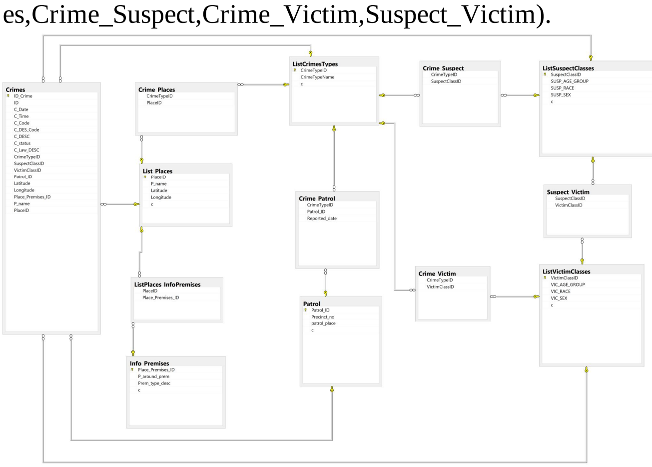


Figure 4: The First Entity Relation Diagram for The Proposed Database

Figure (5) shows the second design of ERD which consist of (Crimes_Full,List_Places,ListSuspectClasses,ListVictimClasses) tables and relations (Crime_Places, Crime_Suspect, Crime_Victim).

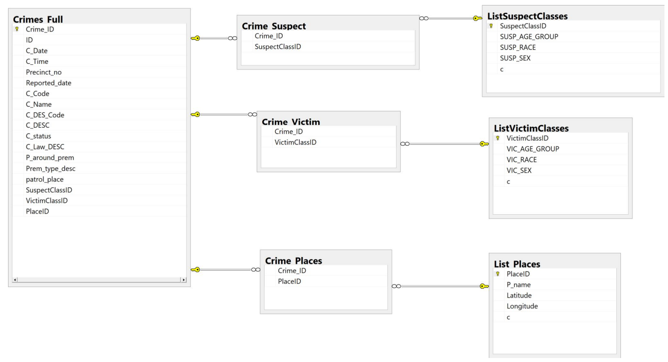


Figure 5: The Second Entity Relation Diagram for The Proposed Database

Figure (6) shows the third design of ERD which consist of (crimes, ListCrimesTypes, ListPlaces, Info_Premises, ListSuspectClasses, ListVictimClasses) tables and relations (Crime_Places,Crime_Place_Victim,Crime_Place_Victim_Suspect,ListPlaces_InfoPremises).

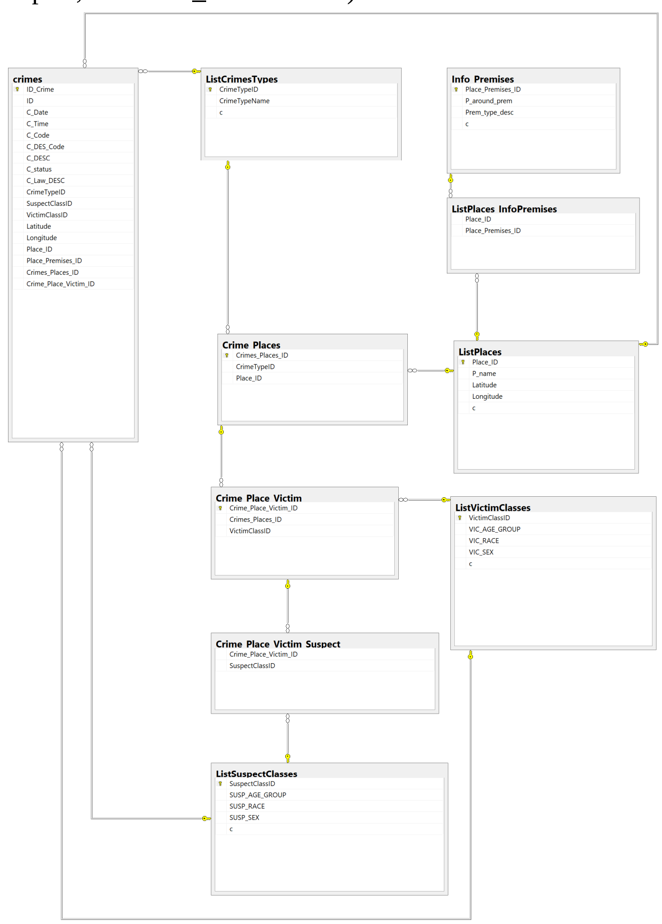


Figure 6: The Third Entity Relation Diagram for The Proposed Database

3.5 Data Warehouse Design

In this stage, a galaxy schema data warehouse design which contains various fact connected with multiple dimension that sorted in one level or more (normalized). The different fact tables and dimensions were selected from the database tables and the required analysis.

3.5.1 Dimensions

The dimensions of the DW has been choosing from the data base, the dimensions called the "soul "of DW. Various analysis can be applied on the DW and thus create a number of dimensions such as the type of crime, the time of the crime, the law description for a crime, the crime location name and description it, the patrol that deals with the crime and some specific information about a crime. Table (3) describe each dimension in the proposed data warehouse.

Table 3: Dimensions Table

	Diminution Table	Descriptive Attributes
1	ListCrimesTypes	CrimeTypeID, CrimeTypeName
2	Crimes	ID_Crime,ID,Crime_Type_ID,C_Time,Precinct_no,Reported_date,C_DES_Code,C_DESC,Crime_Status_ID,Crime_Law_Desc_ID,Crime_date_ID,Date_Year_ID,Date_Month_ID,Date_Day_ID
3	Crime_Date	Crime_date_ID,Crime_Date,Year,Month_Name,Day_Name,Date_Year_ID,Date_Month_ID,Date_Day_ID
4	Crime_Law_Desc	Crime_Law_Desc_ID,Law_Description
5	Crime_Status	Crime_Status_ID, Crime_Status
6	ListPlaces	Place_ID, P_name, Latitude, Longitude
7	ListPlaces_InfoPremises	Place_ID, Place_Premises_ID
8	Info_Premises	Place_Premises_ID,P_around_prem, Prem_type_desc
9	ListSuspectClasses	SuspectClassID,SUSP_AGE_GROUP,SUSP_RACE,SUSP_SeX
10	ListVictimClasses	VictimClassID,VIC_AGE_GROUP, VIC_RACE, VIC_SEX
11	Patrol	Patrol_ID,Precinct_no,patrol_place
12	Dim_Date_Year	Date_Year_ID, Year
13	Dim_Date_Month	Date_Month_ID, Month
14	Dim_Date_Day	Date_Day_ID, Day_Name

3.5.2 Facts

Fact tables represent the place where almost all the analytical processes are applied [27], and hence they contain the dimension key and the measures. In this step of a design data warehouse is to choose carefully, the fact that will appear in the facts table. Can be obtained different reports (analysis). From the fact tables that mentioned in Table (4) such as the crime type with location and suspect, the measure of crime type with location and victims, and the measure of crime type and the Patrol. Each fact table measure (s) is distributed on the timeline (day, month and year).

Table 4: Different DW Fact Tables

	Fact Table	Descriptive
1	Fact_Crime_Place_Suspect_Date_Year	CrimeTypeID, Place_ID, Suspect_ID, Date_Year_ID, Measure
2	Fact_Crime_Place_Victim_Date_Year	CrimeTypeID, Place_ID,Victim_ID, Date_Year_ID, Measure
3	Fact_Crime_Date_Year_Patrol	CrimeTypeID,Date_Year_ID,Patrol_ID, Measure
4	Fact_Crime_Date_Month_Patrol	CrimeTypeID,Date_Month_ID,Patrol_ID, Measure
5	Fact_Crime_Place_Suspect_Date_Month	CrimeTypeID, Place_ID, Suspect_ID, Date_Month_ID, Measure
6	Fact_Crime_Place_Victim_Date_Month	CrimeTypeID, Place_ID, Victim_ID, Date_Month_ID, Measure
7	Fact_Crime_Date_Day_Patrol	CrimeTypeID,Date_Day_ID,Patrol_ID, Measure
8	Fact_Crime_Place_Suspect_Dat	CrimeTypeID, Place_ID, Suspect_ID, Date_Day_ID, Measure

	e_Day	
9	Fact_Crime_Place_Victim_Date_Day	CrimeTypeID, Place_ID, Victim_ID, Date_Day_ID, Measure

3.5.3 Extract, Transform, Load (ETL)

By executing queries in SQL server to extract the data from database and transform the data from tables to in dimensions and facts form and finally loading the data into the data warehouse. Sample of ETL queries is shown in Figure (7).

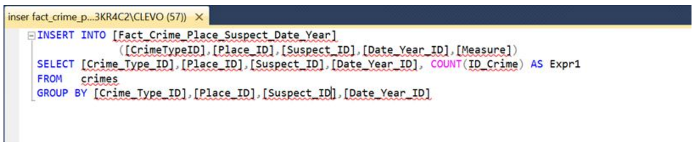


Figure 7: Loading Data into Fact_Crime_Place_Suspect_Date_Year Query

The galaxy schema of the proposed data warehouse is shown in Figure (8).

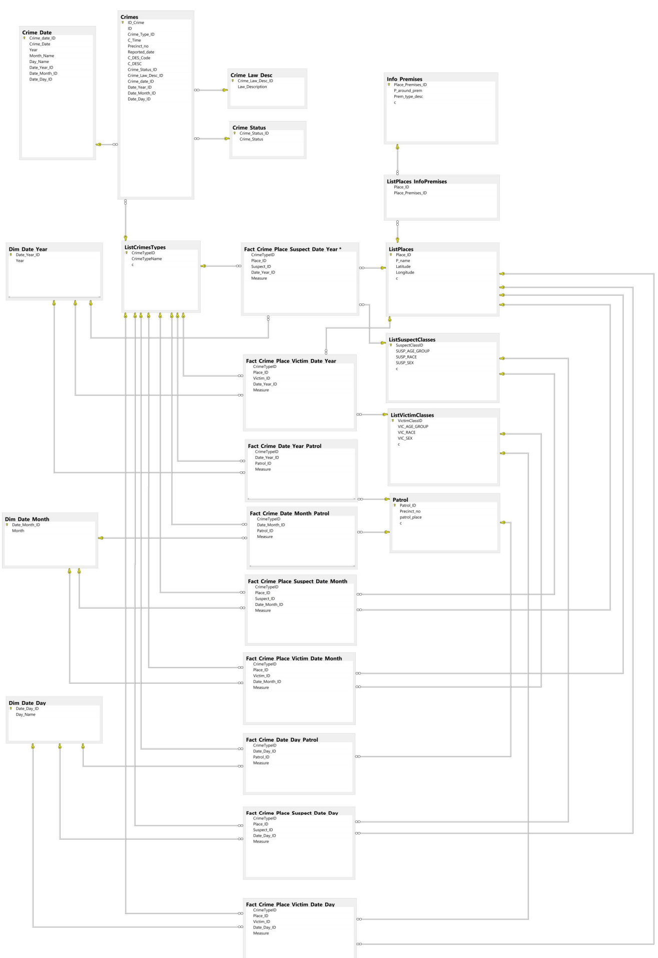


Figure 8: The Proposed Data Warehouse

3.6 Design analysis

The proposed warehouse design fulfils the following issues:

3.6.1 Data Redundancy

In the proposed data warehouse shown in Figure (8), The data redundancy was eliminated such as (X_COORD_CD, Y_COORD_CD, Lat_Lon) columns which were replaced by (Latitude, Longitude) columns. Eliminating the redundancy reduces the data inconsistency, data corruption resulting from errors in writing, reading, storage, or processing data and helping prevent the increase in database size and storage costs.

3.6.2 Access flexibility

Ease of access to data and reports needed by a data analyst, through the facts and dimensions in the proposed data warehouse.

3.6.3 Enforcing security, ownership and privacy

Data security, privacy and ownership on the proposal data warehouse as shown in Figure (10) can be enforced by dividing the data warehouse into multiple data marts each one has fact with its dimensions according to the required reports. Data privacy and ownership was achieved in different data marts to restrict the access of data and reports for authenticated people.

It is important to notice that the NYPD crime dataset doesn't contain some features like job and education for criminal, which helps get more specific analysis to crime and criminal. This problem cannot be solved even using the most advanced techniques in solving missing values and attributes.

4. CONCLUSION AND RECOMMENDATIONS

An optimal galaxy DW for the crime dataset was designed to solve the different features of the analysis including redundancy, access flexibility and enforcing security, ownership and privacy. A structure of data warehouse for crime dataset, as a solution to big data of crime in major cities. We collected real data from New York City (NYPD Complaint Data Historic) available on the internet. The proposed design consists of six stages from collecting data, converting CSV file to SQL table, data preprocessing, design database and design data warehouse finally the analysis. This proposed DW will serve the data analyst and law enforcement forces by getting important analytics, resolve different complex queries as per their need and knowing the complexity of the relationship between crimes with locations, victims and suspects.

It is important to recommend that the proposed data warehouse represent a repository for preprocessed real crime data with huge amount of records (more than six million) available for other research work, and the flexibility of design enables the researcher to add new data and features. Finally, it is of great importance to mention that using a large number of fact tables (nine in this paper) and intern data marts is very important to improve the performance of the analysis algorithms and their complexities.Further, this paper can develop it by using data mining techniques like clustering, classification on the DW to predicted the crime and improve the efforts made by the lawmen and police departments.

REFERENCES

[1] B. B. Miftachul Huda, Andino Maseleno, Pardimin Atmotiyoso, Maragustam Siregar, Roslee Ahmad, Kamarul Azmi Jasmi, Nasrul Hisyam Nor Muhamad, Mohd Ismail Mustari, "Big Data Emerging Technology: Insights into Innovative Environment for Online Learning Resources," *Int. J. Emerg. Technol. Learn.*, vol. 13, no. 1, pp. 23–36, 2018.

[2] M. Feng *et al.*, "Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data," *IEEE Access*, vol. 7. pp. 106111–106123, 2019.

[3] U. Thongsatapornwatana, "A survey of data mining techniques for analyzing crime patterns," *2016 2nd Asian Conference on Defence Technology, ACDT 2016*. pp. 123–128, 2016.

[4] H. K. Fatlawi, A. F. H. Alharan, and N. S. Ali, "An efficient hybrid model for reliable classification of high dimensional data using k-means clustering and bagging ensemble classifier," *J. Theor. Appl. Inf. Technol.*, vol. 96, no. 24, pp. 8379–8398, 2018.

[5] S. H. Farhad Soleimanian Gharehchopogh, "An Optimization K-Modes Clustering Algorithm with Elephant Herding Optimization Algorithm for Crime Clustering," *ournal Adv. Comput. Eng. Technol.*, vol. 6, no. 2, pp. 78–87, 2020.

[6] D. Xu and Y. Tian, "A Comprehensive Survey of Clustering Algorithms," *Annals of Data Science*, vol. 2, no. 2. pp. 165–193, 2015.

[7] S. Hussain, R. Atallah, A. Kamsin, and J. Hazarika, "Classification, clustering and association rule mining in educational datasets using data mining tools: A case study," in *Advances in Intelligent Systems and Computing*, 2019, vol. 765, pp. 196–211.

[8] A. F. H. Alharan, H. K. Fatlawi, and N. S. Ali, "A cluster-based feature selection method for image texture classification," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 3. pp. 1433–1442, 2019.

[9] C. C. Aggarwal and P. S. Yu, "Data mining techniques for associations, clustering and classification," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 1574. pp. 13–23, 1999.

[10] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowl Inf Syst*, vol. 14, pp. 1–37, 2008.

[11] R. Ari Setyawan, E. Prasetyo, and A. S. Girsang, "Design and Implementation Data Warehouse in Insurance Company," *Journal of Physics: Conference Series*, vol. 1175, no. 1. 2019.

[12] M. Subekti, J. Junaidi, H. L. H. S. Warnars, and Y. Heryadi, "The 3 steps of best data warehouse model design with leaning implementation for sales transaction in franchise restaurant," *2017 IEEE International Conference on Cybernetics and Computational Intelligence, CyberneticsCOM 2017 - Proceedings*, vol. 2017-Novem. pp. 170–174, 2018.

[13] C. N. S.Vijayarani, E.Suganya, "A Comprehensive Analysis of Crime Analysis Using Data Mining Techniques," vol. 9, no. 1. pp. 114–123, 2020.

[14] M. C. Giuseppe Agapito, Chiara Zucco, "COVID-WAREHOUSE A Data Warehouse of Italian COVID-19, Pollution, and Climate Data." 2020.

[15] D. B. S. J. CHIRRA ANUSHA, "DESIGN OF A DATA WAREHOUSE FOR MEDICAL INFORMATION SYSTEM USING DATA MINING TECHNIQUES," *J. Eng. Sci.*, vol. 11, no. 6, pp. 308–313, 2020.

[16] A. R. Quitaleg and M. G. Ortiz, "Design and Development of Data Warehouse Framework of Highland Vegetable Crops for Benguet," in *IOP Conference Series: Materials Science and Engineering*, 2020, vol. 803, no. 1.

[17] A. Abdo, H. Fahmy, and A. A. Shaker, "Mining Forensic Medicine Data for Crime Prediction," vol. 17, no. 6. pp. 56–62, 2019.

[18] I. N. A. Prabawa, D. Agung, K. Arimbawa, and I. G. N. Janardana, "Analysis and Design Data Warehouse For E-Travel Business Optimization," *International Journal of Engineering and Emerging Technology*, vol. 4, no. 1. 2019.

[19] Y. P. Sudarmojo, "Design of Library Data Warehouse Using OLTP Result of Services Analysis," vol. 3, no. 1. pp. 62–65, 2018.

[20] N. A. Farooqui and R. Mehra, "Design of a data warehouse for medical information system using data mining techniques," *PDGC 2018 - 2018 5th International Conference on Parallel, Distributed and Grid Computing*. pp. 199–203, 2018.

[21] T. Y. Choo and H. N. Chua, "Data Warehouse Design and Implementation for Research Literature Mining," *Proceedings - 2018 4th International Conference on Advances in Computing, Communication and Automation, ICACCA 2018*. 2018.

[22] I. Sutedja, P. Yudha, N. Khotimah, and C. Vasthi, "Building a Data Warehouse to Support Active Student Management: Analysis and Design," *Proceedings of 2018 International Conference on Information Management and Technology, ICIMTech 2018*. pp. 460–465, 2018.

[23] "NYPD Complaint Data Historic | NYC Open Data." [Online]. Available: https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i. [Accessed: 14-Oct-2020].

[24] A. Kumar and S. P. Panda, "A Survey: How Python Pitches in IT-World," in *Proceedings of the International Conference on Machine Learning, Big*

Data, Cloud and Parallel Computing: Trends, Prespectives and Prospects, COMITCon 2019, 2019, pp. 248–251.

[25] C. Z. & Q. Y. Shichao Zhang, "Data preparation for data mining," *Appl. Artif. Intell.*, pp. 375–381, 2003.

[26] A. Elfaki, A. Aljaedi, and Y. Duan, "Mapping ERD to knowledge graph," in *Proceedings - 2019 IEEE World Congress on Services, SERVICES 2019*, 2019, pp. 110–114.

[27] M. R. Ralph Kimball, *The Data Warehouse Toolkit: The Defi nitive Guide to Dimensional Modeling*, Third Edit. John Wiley & Sons, Inc., 2013.