The Study on Data Warehouse and Data Mining for Birth Registration System of the Surat City

Pushpal Desai
M.Sc.(I.T.) Programme
Veer Narmad South Gujarat University
Surat, India.

Desai Apurva
Department of Computer Science
Veer Narmad South Gujarat University
Surat, India

ABSTRACT
Data Warehousing and Data Mining are widely used by many industries like banking, insurance, healthcare, security and many others, however very little work has been done for e-governance systems in India. The e-governance systems developed by the Surat Municipal Corporation has achieved great success in several years, to service citizens in a more timely, effective, and cost-efficient method. This initiative has resulted in collection of large amount of unexplored and unorganized data. In this paper we proposed Data Warehouse Modeling and Online Analytical Procession (OLAP) for Birth Registration System using Microsoft SQL Server 2008. Our study utilizes data of the Surat city from the year of 2000 to 2009. To query and analyze the data in the data warehouse conveniently and effectively, we designed Data Warehouse using star schema. Our work will help administrators of The Surat Municipal Corporation analyze Birth Registration System data and provide decision-making support for future planning and better service to citizens of the Surat city. Since the research is still in its early stage, the paper mainly focuses on design and implementation of Data Warehouse Modeling, OLAP and Microsoft Data Mining Clustering algorithm for Birth Registration System. The Microsoft Clustering algorithm is used to identify important clusters from the Birth Registration Warehouse.

General Terms
Data Warehouse and Data Mining.

Keywords
OLAP, Clustering, Birth Registration System.

1. INTRODUCTION
In the scenario of ever-changing social and business conditions, organization's needs to have access to more and better information. Almost all organizations are now days using computerized systems as the backbone of their operations but the fact is that despite having a large number of powerful computerized systems and a fast and reliable network, access to information that is already available within the organization is very difficult to access. After implementing several computerized systems typically organization's data are scattered across various databases, flat files, physical records store at different geographical locations. All organizations that use computerized systems for their different operations produce large amount of data. Most of the time this data remains in the operational systems, flat files and can't be used by the organization. In this condition only a small portion of this data that is entered, processed and stored is actually available to decision makers. The unavailability of crucial data can cause significant reduction in efficiency of organizations. Many large organizations found themselves with data scattered across multiple platforms and variations of technology, making it almost impossible for any one individual to use data from multiple sources. It is therefore crucial that organizations have a good source of organization's data that can be used quickly and flexibly by management. Data Warehousing has emerged as a key technology for enterprises that wish to improve their data analysis, decision support activities, and the automatic extraction of knowledge from data. Furthermore Data Mining algorithms can answer organization questions that traditionally are time consuming to resolve. In general, wherever data exist in, powerful data warehouse and data mining techniques can help reveal important data patterns that would otherwise remain unnoticed when using simple type of analysis. Pragmatic use of data warehousing and data mining technologies can contribute a lot to decision support systems in industry automation, E-governance, health care and in many other areas.

2. DATA WAREHOUSE
Data Warehouse has been defined in a different ways by various authors. Inmon defines Data Warehouse as "a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision" [1]. Here "subject-oriented" means the information in the data warehouse is organized according to subject and provides information for subject-oriented decision making process. The meaning of "integrated" is that the information in the data warehouse is not simply extracted from various operation systems but systematically processed, summarized and reduced, which ensure that the information in data warehouse is consistent and comprehensive for a organization. The "stable" means that, generally speaking, once a piece of data is put in the data warehouse, it will be kept permanently. The "time-varying" means that information in the data warehouse is not only on the present situation or a certain time point's situation, but also on situation of various stages from a former time point to present, on which quantitative analysis and forecasting can be made on development history and trend. Fox provides another perspective to define the data warehouse: "Data Warehouse is the process by which an organization sets up and maintains a central repository for significant portions of its transaction processing or program management data, which can be selective, extracted and organized for analytical applications, user queries and report generation" [2]. Data Warehouses are

databases used for storing large amounts of data, collected from multiple data sources. This data is used in knowledge retrieval processes, business intelligence applications, data mining etc. as the organization's primary source of decision making data.

3. DATA MINING

There is no universal agreement towards the definition of Data Mining. Data mining model integrates various techniques and fields, it has meant different things to different group of people and hence many authors have given definitions in their own way. According to Fayyad "Data mining is a step in the knowledge discovery in databases (KDD) process and refers to algorithms that are applied to extract patterns from the data. The extracted information can then be used to form a prediction or classification model, identify trends and associations, refine an existing model, or provide a summary of the database being mined" [4]. Brabazon describes Data mining is the discovery of new, non-obvious, valuable information from a large collection of raw data [5]. In recent past Liao describes "Data Mining (DM) is an interdisciplinary field that combines artificial intelligence, computer science, machine learning, database management, data visualization, mathematic algorithms, and statistics. DM is a technology for knowledge discovery in databases (KDD). This technology provides different methodologies for decision making, problem solving, analysis, planning, diagnosis, detection, integration, prevention, learning and innovation." [7].

4. THE PROBLEM

In recent years governments all over the world have been successful in implementing various E-governance projects. E-governance allows government to service their citizen in a better way. Not only E-governance provide more facilities to citizen but also make government accountable which is crucial in developing nation like India. The E-governance projects implemented in the Surat city have already paid rich dividends for Surat Municipal Corporation and citizens of the city. Surat Municipal Corporation has taken early initiatives in E-governance and hence different manual tasks are computerized and processes are simplified. This has resulted in timely, effective and cost-effective service to citizen of Surat city. Currently many processes like Child Birth registration, Death registration, Vehicle registration, Property registration etc…are fully computerized and data are stored in centralized database system. This automation has resulted in gigabytes of data containing millions of records regarding various aspects of population of the Surat city. All this information is scattered and maintained in different format. Figure 1 describes current state of information scattered across various databases, files, reports and websites. Since last few years many government and non-government organizations have invested heavily for computerization of various processes. As various processes are computerized, that has resulted in collection of large amount of data residing in different operational systems. All this data most of the time remains idle and not utilized. By building Data Warehouse on huge amount of data already available, there is good potential of knowledge discovery that can be utilized by different organizations and our society. In this research paper we have focused on design and implementation of Data Warehouse and Data Mining's Clustering algorithm for Birth Registration data.

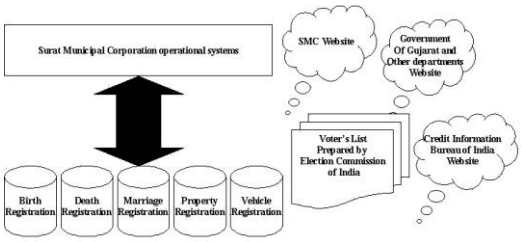


Figure. 1. Current State of Information at Surat Municipal Corporation

5. RESEARCH METHODOLOGY

We used Microsoft SQL Server 2008 Analysis Service to build Data Warehouse for Birth Registration System. The process of extracting data from Birth Registration Systems and loading into Data Warehouse is commonly known as ETL, which stands for extraction, transformation, and loading. To implement ETL process Microsoft SQL Server 2008 was used and Data Warehouse was established using Microsoft Analysis Services. We used Microsoft Clustering algorithm for identifying important clusters in the Birth Registration data for the city of Surat.

5.1 Designing of conceptual model

At present, there are two commonly used conceptual models in Data Warehouse: Star Model and Snowflake Model. We have adopted Star model in the Data Warehouse system for reducing scanning time in the fact tables and improving capability of inquiring. Following table contains list of dimensions which connects to fact table:

Table 1. The Dimensions of The Birth Registration System

Subject	**Dimension ID**
Birth Location	BirthLocation ID
Year	YearID
Religion	ReligionID
Gender	GenderID
Education	EducationID
Zone	ZoneID
Delivery Attention	DeliveryAttentionID
Delivery Method	DeliveryMethodID

5.2 Designing of logical model

After analysing various dimensions and subjects for Data Warehouse, we developed Data Warehouse using Star Model. Following diagram shows structure of the star model. The Birth Data fact table is at the center multidimensional Data Warehouse. The fact table contains information about Birth Registration and different keys that connects fact table with various dimension tables. Dimensions are stored in dimension tables that contain dimensional elements and attributes.

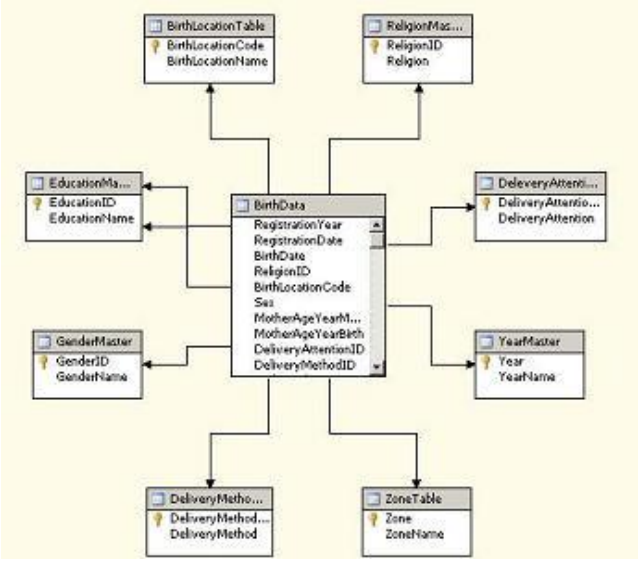


Figure 2. Star Schema of Birth Registration Data

5.3 Results

Online Analytical Processing is extremely useful for multidimensional data analysis. The objective of using OLAP is to help decision maker utilize data and information effectively. Multidimensional data set is the core mechanism in OLAP for data analysis. Analysis Services of SQL Server 2008, allows us to Slice, Dice, Rotate and Drill that converts data into information. As shown in Figure 2 and Figure 3, OLAP technology is used to construct multidimensional data set for Birth Registration data.

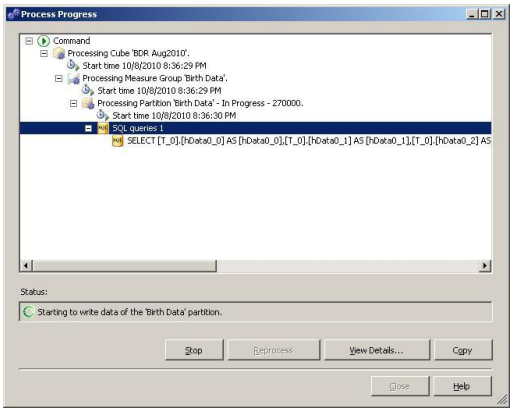
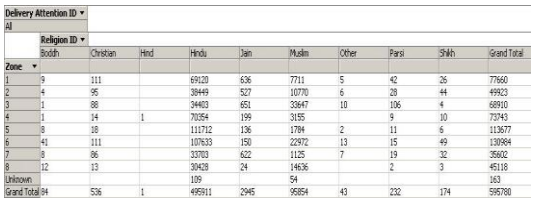


Figure.3. Processing of multidimensional data

Depending upon the organization requirements fact measures can be selected from the fact table. For example, we have created Birth Registration Cube having Delivery Method, Zone and Religion as three different dimensions. For Cube design, Multidimensional Online Analytical Processing (MOLAP) data storage was chosen and data can be traversed from any arbitrary angle by browse option. Following figure shows browse data with various dimensions.



Delivery Attention ID: All										
	Religion ID	Boddh	Christian	Hind	Hindu	Jain	Muslim	Other	Parsi	Shikh	Grand Total
Zone											
1		9	111		69120	636	7711	5	42	26	77660
2		4	95		38449	527	10770	6	28	44	49923
3		1	88		34403	651	33647	10	106	4	68910
4		1	14	1	70054	199	3155		9	10	73743
5		8	18		111712	136	1784	2	11	6	113677
6		41	111		107633	150	22972	13	15	49	130984
7		8	86		33703	622	1125	7	19	32	35602
8		12	13		30428	24	14636		2	3	45118
Unknown					109		54				163
Grand Total		84	536	1	495911	2945	95854	43	232	174	595780

Figure 4. 3-Dimensional (Delivery Method, Zone and Religion) data browse of Birth Registration Data.

Here, it is possible to increase or decrease dimension levels and angles. We can also form combinations of different dimensions from existing dimension tables. In another Cube, we considered four dimensions – Birth Location, Gender, Zone and Year for Birth Registration data. Following figure shows browse data considering four dimensions.

Birth Location Code: All		Gender ID		
		Female	Male	Grand Total
Zone	Year	Birth Data Count	Birth Data Count	Birth Data Count
⊞ 1		34832	42828	77660
⊞ 2		22856	27067	49923
⊞ 3		32268	36642	68910
⊞ 4		31604	42139	73743
⊞ 5		46751	66926	113677
⊞ 6		58401	72583	130984
⊞ 7		16225	19377	35602
⊞ 8		20833	24285	45118
⊞ Unknown		83	80	163
Grand Total		263853	331927	595780

Figure 5. 4-Dimensional (Birth Location, Zone, Year and Gender) data browse of Birth Registration Data.

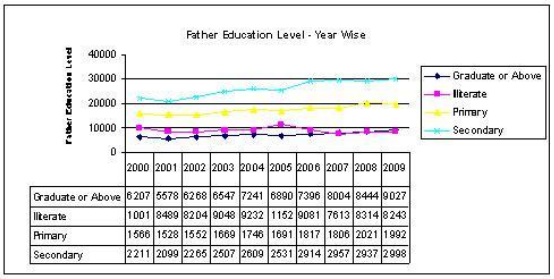
Once we create Cubes in Data Warehouse it is extremely easy for end users to explore data by using different perspectives. In the above-mentioned cube, we can drill data Zone wise, Year Wise, Gender wise and/or Birth location wise. Following figure shows Data Drilling concept where Zone equal to 1, Female & Male Birth Data Count, Year and Birth Location Code.

Birth Location Code: (Multiple Items)		Gender ID		
		Female	Male	Grand Total
Zone	Year	Birth Data Count	Birth Data Count	Birth Data Count
⊟ 1	2000	3691	4435	8126
	2001	3280	4107	7387
	2002	3382	4449	7831
	2003	4083	5093	9176
	2004	4221	5367	9588
	2005	4281	5508	9789
	2006	3304	3910	7214
	2007	2625	3110	5735
	2008	2784	3182	5966
	2009	3041	3487	6528
	Total	34692	42648	77340

Figure 6. Data drilling.

In another cube, we explored the relationship between Birth Registration and Parents Education Level. We observed that father education information was provided for all records. However, mother education information was provided only in 1,43,604 records out of 5,95,780. For remaining 4,52,176 records mother education information field was blank. The parents education level can be categorised into for levels: Graduate or Above, Illiterate, Primary, and Secondary. Based on these categories and Birth Registration data we developed cube that

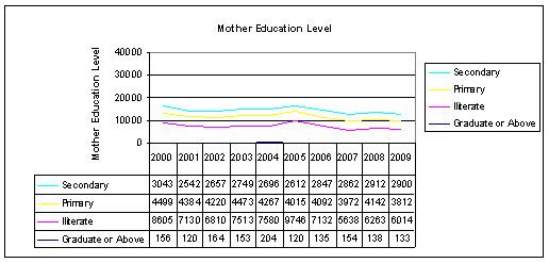
show relationship between father education level and birth registration year wise.



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Graduate or Above	6207	5578	6268	6547	7241	6890	7396	8004	8444	9027
Illiterate	1001	8489	8204	9048	9232	1152	9081	7613	8314	8243
Primary	1566	1528	1552	1669	1746	1691	1817	1806	2021	1992
Secondary	2211	2099	2265	2507	2609	2531	2914	2957	2937	2998

Figure 7. Chart 1. 2-Dimensional (Father Education and Year) data browse with chart.

Similarly, chart was prepared for mother education level and birth registration. To obtain desired result two measures Father Education Count and Mother Education Count were used in the cube.



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Secondary	3043	2542	2657	2749	2696	2612	2847	2862	2912	2900
Primary	4499	4384	4220	4473	4267	4015	4092	3972	4142	3812
Illiterate	8605	7130	6810	7513	7580	9746	7132	5638	6263	6014
Graduate or Above	156	120	164	153	204	120	135	154	138	133

Figure 8. Chart 2. 2-Dimensional (Mother Education and Year) data browse with chart.

We used Microsoft Clustering Algorithm to identify important clusters in the Birth Registration Data. Figure 9. Describes relationship among important clusters in the Birth Registration data.

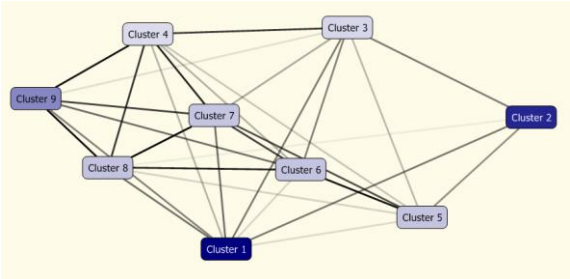


Figure 9. Cluster information for Birth Registration Data.

Microsoft Clustering Algorithm also allows you to explore profile of every Cluster generated via algorithm. Figure 10. describes cluster profile with various attributes and their states.

Figure 10. Cluster Profile for Birth Registration Data.

6. CONCLUSION

This paper gives an insight on Data Warehouse modeling, OLAP and Data Mining Clustering algorithm for Birth Registration system, which will provide the effective support to the Municipal Corporation. In this paper we have used Microsoft Clustering algorithm to identified important clusters that demonstrates relationship between Religions, Sex, Delivery Methods and Delivery Attention. Data Mining Cluster algorithm profile allows administrator of the Surat Municipal Corporation to identify useful clusters, which can help them in better planning and decision-making. The traditional system can only collect data but does not provide knowledge of strategic significance to administrators and decision makers of the Municipal Corporation. The main advantage of building Data Warehouse is that administrators of the Municipal Corporation can view data from different perspectives, define various metrics of interest and query data at any level of detail using various methods such as Slice, Dice, Rotate and Drill.

7. REFERENCES

[1] William H.Inmon, Building the Data Warehouse. 2006. Fourth Edition,New York:John Wiley & Sons.

[2] Fox. 2000. A.: Data warehousing: avoiding the pitfalls, Behavioral Health Management, Vol. 20, No. 3, p. 18.

[3] Li Tong Cui Yan Ren Shu Po. 2008. Analysis on Data Warehouse Technology and Its Development Situation, IEEE International Symposium on Information Science and Engieering, pp. 486--488.

[4] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. 1996 The KDD process for extracting useful knowledge from volumes of data, Association for Computing Machinery. Communications of the ACM, 39(11), 27--34.

[5] Brabazon, T. 1997. Data mining: A new source of competitive advantage? Accountancy Ireland, pp 30--31.

[6] Li Tong, Cui Yan et al. 2008. Analysis on Data Warehouse Technology and Its Development Situtation, IEEE International Symposium on Information Science and Engineering, pp. 486-488.

[7] Liao, S. h. 2003. Knowledge Management Technologies and applications-Literature review from 1995 to 2002. Expert System with Application 25, Pergamon, pp. 155--164.

[8] Thomas Neumuth, Svetlana Mansmann et al. 2008. Data Warehousing Technology for Surgical Workflow Analysis,

21st IEEE International Symposium on Computer-Based Medical Systems, pp. 230--235.

[9] Wang Kuanfu, Hu Xuanzi. 2008. Application of Data Warehouse Technology in Data Center Design, IEEE International Conference on Computational Intelligence and Security, pp. 484-488.

[10] Wen, C. P. 2004. Hierarchical Analysis For Discovering Knowledge in Large Databases, Information Systems management, pp. 81—88.

[11] Xiaofeng Zhang. 2008. A New Modelling Method for the Data Analysis Solution in Business. IEEE International Symposium on Electronics Commerce and Security, pp. 175—178.

[12] Xuezhong, Baoyan et al. 2008. Building Clinical Data Warehouse for Traditional Chinese Medical Knowladge Discovery, IEEE International Conference on Bio-Medical Engineering and Informatics, pp. 615--620.

[13] ZHANG Dan-Ping. 2009. A Data Warehouse Based on University Human Resources Management of Performance Evaluation, IEEE International Forum on Information Technology and Application, pp. 655—658.

[14] ZHOU Qian, XIAO Qing. 2009. The Study on Data Warehouse Modelling and OLAP for Highway Management, IEEE International Conference on Measuring Technology and Mechatronics Automation, pp. 416--419.

[15] Desai Pushpal, Desai Appurva, 2011. The Study on Data Warehouse Modelling and OLAP for Birth Registration System of the Surat City. TECHNOLOGY SYSTEMS AND MANAGEMENT, Communications in Computer and Information Science, 2011, Volume 145, Part 1, 160-167.