Multidimensional Analysis for Census Data by Applying Star Schema Model

Myint Myint Thein, Myint Myint Lwin, Aye Chan Mon, May Thu Aung
University of Information Technology,
University of Computer Studies (Maubin), Myanmar
{myintmyintthein, myintmyintlwin, ayechanmon, maythuaung}@uit.edu.mm

Abstract

In recent years, the high value of multidimensional data has been recognized in both the academic and business communities. Star schemas are the primary storage mechanism for multidimensional data that is to be queried efficiently. It supports relationships between fact and dimension tables and creating combination dimensions with a key, resulted to improve query performance for large quantities of data. This paper is presented for multidimensional data model, that is called star schema to store large amount of census data. This star schema can be used for business related queries on Census data for visualization report. This paper aims to enhance the interactive visualization process with more relevant operations for manipulation of various attributes by using the Pentaho Business Analytics (BA) Suite.

Keywords—Agglomerative, census data, clustering, K-means

1. Introduction

In recent years, large multidimensional databases or data warehouses have become common in a variety of applications. Data warehouse (DW) is a collection of technologies that is enabling the decision maker to make better and faster decisions. The major challenge with these databases is to extract meaning from the data, they contain discover structure, find patterns and derive causal relationships. Collecting data for a Business Intelligence (BI) application is done by building a data warehouse where data from multiple heterogeneous data sources is stored. Transferring the data from the data sources to the data warehouse is often referred to as the Extract, Transform and Load (ETL) process. The data is extracted from the source, transformed to fit and finally the data is loaded into the warehouse. The ETL process often brings issues with data consistency between data sources. In order to load it into the data warehouse the data has to be consistent, and the process to accomplish this is called data cleaning.

A star schema is a method of organizing information in a data warehouse that enables efficient retrieval of business information. The star schema represents to the end user a simple and query-centric view of the data by partitioning the data into two groups of tables: facts and dimensions. Facts are the data keys being organized around a large central table, and dimensions contain the metadata that is related to a set of typically smaller tables. Data stored in a star schema is defined as being "denormalized." Denormalized means that the data has been efficiently structured for reporting purposes. The goal of the star schema is to maintain enough information in the fact table and related dimension tables so that no more than one join level is required to answer most business-related queries.

This paper represents star schema model for storing Census data. The data is retrieved from data sources by making ETL process in Pentaho Data Integration (PDI) open source tool and loaded into the star schema data warehouse on PostgreSQL database. It provides the useful information to the appropriate decision makers for business decision.

2. Related Work

Sudhir B. Jagtap and Kodge B.G. [4] have made an attempt to demonstrate how one can extract the local (district) level census, socio-economic and population related other data for knowledge discovery and their analysis using the powerful data mining tool Weka. Their primary available data such as census (2001), socio-economic data, and

ISBN 978-99971-0-381-9 © 2017 ICAIT

few basic informa- tion of Latur district are collected from National Informatics Centre (NIC), Latur, which is mainly required to design and develop the database for Latur district of Maharashtra state of India. The database is designed in MS-Access 2003 database management system to store the collected data and analyzed data by using Weka tool.

M. Yost, J. Nealon presented a star schema model [1] that can be used for any census or survey to track the full history of the data series and to standardize the metadata. The dimensional model represents a relational database model that facilitates the gathering of a great deal of this information and knowledge about the data, stores it, organizes it and then relates it directly to the factual data being analyzed.

O.C. Okeke, B.C. Ekechukwu [3] proposed to develop mining model applicable to the analysis of Nigeria census data by harnessing the power of data-mining technique that could uncover some hidden patterns to get their geo-spatial distribution. This is represented decision tree learning for approximating discrete-valued target function and decision tree algorithm was used to predict some basic attributes of population in the census database.

N. Stolba, A.M. Tjoa [2] to explain the integration of data warehousing, OLAP and data mining techniques in the field of health care, and an easy to use decision support platform, which supports the decision making process is the process of building care providers and clinical directors. They offered three case studies, which show that the clinical data warehouse that facilitates evidence-based medicine is a platform for reliable, robust and easy to use to make strategic decisions, which are of great importance to the practice of medicine, and the acceptance of evidence-based.

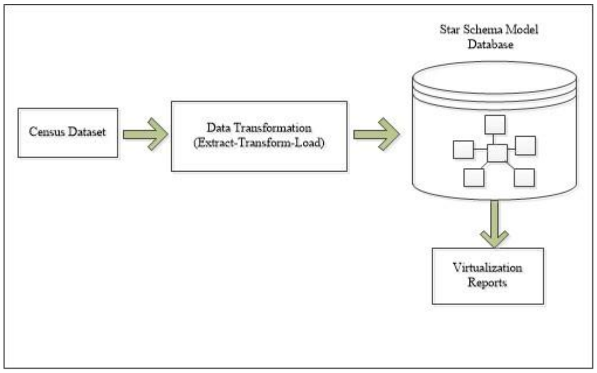

Fig. 1: Proposed System Flow

M.Khan and S.S.Khan [7] surveyed about various visu- alization methods. The authors point out different available visualization techniques are used for different situation and data mining techniques, mining results can present effectively by using visualization methods.

3. Proposed System Architecture

The 2014 Myanmar Population and Housing Census data was undertaken by the Ministry of Immigration and Population that is a fundamental source of information, it includes main categories such as education, fertility, mortality, migration, disability, population projections, gender, housing conditions and assets, youth, and elderly. In Fig. 1, this proposed system illustrates the transformation from dataset to product analysis reports by creating multi-dimensional model (star schema). The first stage started with data preparation for real census data which may involve cleaning data and selecting subsets of records related with education category. The second stage used data transformation (Extraction-Transformation-Loading ETL) tool is data extraction from raw data file (.excel) and then insertion into fact and dimension tables on the PostgreSQL database. The dimension tables are age, gender, literacy, attendance, grade and employment and the one fact table is education table. The final stage demonstrated the impact of age, literacy, attendance, grade and employment on the gender class to produce analysis reports by using Pentaho dashboard tool.

A. Star Schema Model

The relational model and the Structured Query Language (SQL) allows the user to efficiently and effectively manipulate a database. They record information in two dimensions table structure and automate repetitive tasks. Simple entity relationship diagram (ER) consists of many relationships between tables to retrieve query. For this retrieve query, the user needs to fulfill certain criteria. If the user could be required query result, the query operation needs at least one internal join across two tables. Thus, it could be acceded lower performance significantly because of the time required for joining many tables. [6] Another problem is that large amount of data in form of normalization requires a lot joins of many tables. If the user used this model for analysis, it responds very slowly to new analytical requirements.

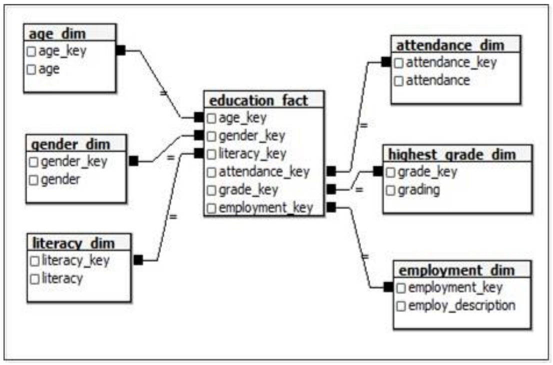

Fig. 2: Star Schema Design

Data warehousing provides an effective way to analyze the data and also uses query optimization by creating multi- dimensional data tables are called star schema. The star schema represents to the end user a simple and query-centric view of the data by partitioning the data into two groups of tables: facts and dimensions. Facts are the data keys being tracked, and dimensions contain the metadata describing the facts. A star schema consists of a collection of tables that are logically related to each other.

In Fig.2, shows a star schema organized our education information from census data, there are one education fact and six main dimensions: age, gender, literacy, attendance, grade and employment. The fact table relates data keys to each dimension. Each dimension tables contain detail information on persons responding to census data. For example, the age tables defines different age level on different persons.

B. Data Transformation Tool

The ETL tool has three tasks to build data warehouse: (1) data is extracted from different data sources, (2) propagated to the data staging area where it is transformed and cleansed, and then (3) loaded to the data warehouse [5]. An ETL system consists of three functional steps:

Extraction: This step is responsible for extracting data from the source systems. Each data source has its distinct set of characteristics that need to be managed in order to effectively extract data for the ETL process.

Transformation: This step tends to make the extracted data to gain accurate data which is correct, complete, consis- tent, and unambiguous. This process includes data cleaning, transformation, and integration. It defines the granularity of fact tables, the dimension tables, data warehouse (star or snowflake), derived facts, and slowly changing dimensions.

Loading: In this loading step, extracted and transformed data is written into the dimensional structures actually accessed by the end users and application systems. Loading step includes both loading dimension tables and fact tables.

In this data transformation step, we used Pentaho Data Integration tool that is free open source data warehousing tool for academic and research purpose. This tool provides access to PostgreSQL database directly. In Fig.3, we extracted data from census data file (excel), the related data inserted into each of age, literacy, attendance, grade, employment and gender dimension tables and education fact table.

The education fact table is 10143 instances. The number of records of dimension tables will be shown in Data Description Section. By applying this tool, the total transformation time of data processing is 10.8 msec.

Table 1 (a) Education Fact Table

age	literacy	attendance	grade	employment	gender
1	1	1	1	1	1
2	2	2	2	2	2
3	2	2	2	2	2
4	1	1	1	3	1
5	1	1	3	3	3
...	...	...	...	...	...
...	...	...	...	...	...

Table 2 (b) Grading

grade-key	grading
1	NG
2	none
3	other
4	1
5	2
6	3
7	4
8	5
9	6
10	7
11	8
12	9
13	10
14	11
15	college
16	vocationaltraining
17	undergraduate
18	graduate
19	master
20	Ph.d
21	postdoc

Table 3 (c) Age

age-key	age-range
1	0-10
2	11-20
3	21-30
4	31-40
5	41-50
6	51-60
7	61-70
8	71-80
9	81-90
10	91-100

Table 4 (d) Attendance

attend-key	attendance
1	previously
2	currently
3	NG

Table 5 (e) Employ

employ-key	employ
1	NG
2	other
3	student
4	gov-employer
5	org-employer
6	own-worker
7	family-worker
8	household-
9	unexper-worker
10	withexper-
11	general-worker
12	retired
13	disabled

Table 6 (f) Literacy

literacy-key	literacy
1	yes
2	NG
3	no

Table 7 (g) Gender

gender-key	gender
1	male
2	NG
3	female

C. Dataset Description

TABLE I. shows the database is designed in PostgreSQL9.5 database management system to store the collected data. In database, there are one education fact and six dimension tables: age, gender, literacy, attendance, grading and employment. The fact table relates data keys to each dimension. Each dimension tables contain detail information on persons responding to real-world dataset. The education fact table has 6 attributes and 10143 records (instance) in the dataset. For example TABLE IV: (d) shows the attendance dimension table contains previously, currently and NG they are not given. TABLE VII: (g) shows the gender dimension table contains male, female and NULL they are not given.

4. Data Visualization

A. Visualization

Data are presented in a variety of formats and it is difficult to analyze when it is needed to use for decision making pro- cesses. Visualization is the best solution to produce graphical representations of data or concepts for decision making. Data visualization can provide large or complex data with a well-designed visual or graphics and can add value for underlying data. Large data are difficult to identify without the help of visuals because most people cannot interpret a table of data without time consuming analysis and tables of data cannot make a decision efficiently. Visuals can turn a data table into a graphic that can be quickly interpreted so it is easier to use evaluation findings. Visualization methods can visualize the innumerate amount of the analytical results as diagrams, tables and images. Visualization is the best solution to produce graphical representations of large amount of data or concepts for decision making. Visualization for Big Data differs from all of the previously traditional visualization techniques. A visual can be either static or interactive. Static visuals are the most common since they are the simplest to produce. Interactive data visualization is a technique of analyzing data, where a user interacts with the system that results in visual patterns for a given set of data.

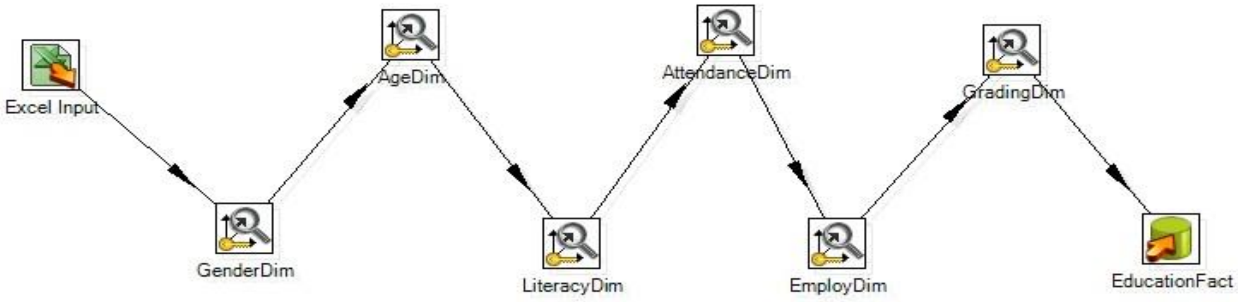

Fig. 3: Data Transformation Step

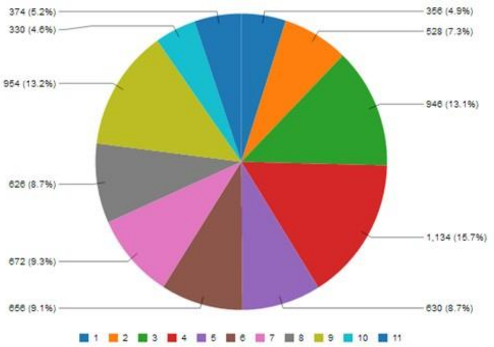

Fig. 4: Analysis for Grading with Male

B. Visualizations for Census Data

Data visualization is using techniques that extract useful information. Visualization will make it easier to extract busi- ness intelligence. The ability to make timely decisions based on census data is crucial to decision making in national development. To display interactive visualization process, the visualization tools Pentaho Business Analytics (BA) Suite generated census data according to various attributes. This tool is a platform which can access, integrate, manipulate, visualize, and analyze your data. Whether data is stored in a flat file, relational database, Hadoop, NoSQL database, analytic database, social media streams, operational stores, or in the cloud, this tool can discover, analyze, and visualize data. The Pentaho will create advanced visualizations of data and provide powerful insight of data.

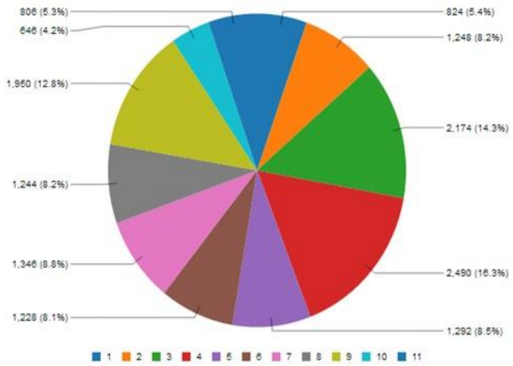

Fig. 5: Analysis for Grading with Male and Female

C. Data Visualizations Results

Data visualization is the study of representing data in some systematic form, including attributes and variables for the unit of information. These data presentation should be beautiful, elegant, descriptive, and interpretable in order to convey message to the reader effectively. Data visualization represents data in the way that simplifies data interpretation and its relationship. The following section is a summary of the data used in the data collection and creation of the visualizations for Myanmar census data.

In Fig.4, shows that the analysis of primary to high school students for Male. The children is age 4 year as kindergarten, they are attending at preschool, the percentage of preschool children are 5.4 and number of children is 824. The number of primary one children is 1248 and 8.2%. The number of primary two children is 2174 and 14.3%. The largest percentage is 16.3% and the number of primary three children is 2490. The number of primary four children is 1292 and 8.5%. The number of primary five children is 1228 and 8.1%. The number of primary six student in high school is 1346 and 8.8%. The number of primary seven student is 1244 and 8.2%. The number of primary eight student is 1950 and 12.8%. The number of primary nine student is 646 and 4.2%. The number of primary ten student is 806 and 5.3%, they are matriculation for attending university.

In Fig.5 describes that the analysis of male and female between primary to high school grading. The largest number of male and female is 2490 and the total percentage is 16.3 of primary three children. The smallest of male and female in primary nine students is 646 and 4.2%. In this paper, our visualization reports are not enough to leverage the benefits of business intelligence in such a dynamic industry.

5. Conclusion

This paper represented on a case study in which data warehouse tool has been applied to some data that is a portion of real census data. And then, it focused on the database model is star schema and converted data into the postgreSQL database. The star join schema represents a relational database model that facilitates the gathering of a great deal of this information and knowledge about the data, stores it, organizes it, and then relates it directly to the factual data being analyzed. Data-mining helps governments, individuals, companies to uncover hidden patterns in large database which is used for development and making decision from virtualization reports using census data. For future work, we have to present the needed and right information should also find a way to get to the right people in right time. We have to do these reports not only help to understand the past, but also work to find new opportunities and emerging trends in future.

6. References

[1] M. Yost and J. Nealon, Using A Dimensional Data warehouse to standardize survey and Census Metadata.

[2] N. Stolba and A.M. Tjoa, The relevance of data warehousing and data mining in the field of evidence based medicine to support healthcare decision making. European Social Fund (ESF), under grant 31.963/46- VII/9.

[3] O.C. Okeke and B.C. Ekechukwu, Using Data-Mining Technique for Cen- sus Analysis to Give Geo-Spatial Distribution of Nigeria. IOSR Journal of Computer Engineering (IOSR-JCE), p-ISSN: 2278-872 Volume 14, Issue 2(Sep-Oct, 2013),PP01-05.

[4] S.B. Jagtap and Kodge B.G, Census Data Mining and Data Analysis using WEKA. International Conference in ”Emerging Trends in Science, Technology and Management-2013, Singapore.

[5] S.H. Ali El-Sappagh, A.M. Ahmed Hendawi and A.H. El Bastawissy, A proposed model for data warehouse ETL processes, Journal of King Saud University– Computer and Information Sciences (2011) 23,91–104

[6] W.Q. Qwaider, Apply On-Line Analytical Processing (OLAP) with Data Mining For Clinical Decision Support International Journal of Managing Information Technology (IJMIT), Vol.4, No.1, February 2012

[7] M. Khan and S.S.Khanr, Data and Information Visualization Methods, and Interactive Mechanisms: A Survey, International Journal of Computer Applications (0975 – 8887), Volume 34– No.1, November 20.