Design and Implementation of Data Warehouse with Data Model using Survey-based Services Data

Boon Keong Seah
MIMOS
Technology Park Malaysia, Malaysia
bk.seah@mimos.my

Nor Ezam Selan
MIMOS
Technology Park Malaysia, Malaysia
nor.ezam@mimos.my

***Abstract*—Various business organization or government bodies are enhancing their decision making capabilities using data warehouse. For government bodies, data warehouse provides a means by enabling policy making to be formulated much easier based on available data such as survey-based services data. In this paper we present a survey-based service data with the design and implementation of a Data Warehouse framework for data mining and business intelligence reporting. In the design of the data warehouse, we developed a multi-dimensional Data Model for the creation of multiple data marts and design of an ETL process for populating the data marts from the data source. The development of multiple data marts will enable easier report generation by identifying common dimension amongst the data marts. The cross-join capabilities of the data marts through common dimensions, demonstrate the ability to easily drill across the data marts for cross data analysis and reporting. In addition, we also have incorporate data quality checking on the data source as well as data detection rules to filter out unmatched data schema and data range from being stored in the data warehouse for analysis.**

Keywords—data model, data warehouse; Extract Transform Load (ETL); OLAP; business intelligence, data mining, star schema, data marts

I. INTRODUCTION

In order to better provide an environment for government, organisations as well as business community in planning and decision making, survey-based services data are collected from various industries to help forms policies decisions. Nevertheless, to achieve this there is a need for the implementation of a business intelligence dashboard and data mining which relies heavily on the formulation of the data warehouse and the data model for enabling such activities. In this paper we present a survey-based service analysis with the design and implementation of a Data Warehouse framework for data mining and business intelligence reporting. In the design of the data warehouse, we developed a multi-dimensional Data Model for the creation of multiple data marts for the data analysis. With the multiple data marts, it is easier to cater for each report needs by identifying common dimension amongst the data marts. The cross-join capabilities of the data marts through common dimensions, demonstrate the ability to easily drill across the data marts. In addition, we also designed an ETL process for data population from data source to the data warehouse. The construction of the system had two stages, the first stage was the data model design and data loading, the second stage was application development that based on the data warehouse. At present, application based on the data warehouse mainly have two types, they were pre-defined reports and OLAP analysis, with addition of data mining supplemented to meet requirements of higher management.

In this paper, we detailed in Section II the background of the principle of having the data warehouse. Section III gives an overview of the data model and data warehouse design while Section IV presents the data model consideration and Section V details the data model implemented. In Section VI, we describe the ETL design process implemented to extract the data source until the loading to the data warehouse. In Section VII, we present the result of the data model and data warehouse and finally in Section VIII, we present the conclusion of this paper.

II. DATA WAREHOUSE BACKGROUND

Data warehouse terminology by Ralph Kimball, defines data warehouse as "a copy of transaction data specifically structured for query and analysis" [9]. Data warehouse of an enterprise consolidates data from heterogeneous sources to support enterprise wide decision-making, reporting, and analyzing. Data warehouses often use star and snowflake schemas [14] to provide the fastest possible response times to complex queries. Bill Inmon [6] proposed the snowflake schema which is a variant of the star schema model, where some dimension tables are normalised, thereby further splitting the data into additional tables. Ralph Kimball [3, 4, 5] proposed the star schema for representing multidimensional data where the schema graph resembles a starburst, with the dimension tables displayed in a radial pattern around the central fact table.

In general, populating any data warehouse consists of the following steps:

A. Data Loading

Data consolidated from heterogeneous data sources may have problems, and needs to be first transformed and cleaned before loading into the DW. The data may have incorrect data value such as null value, inconsistent reference code, and others. Hence, data cleansing is an essential task in data warehousing process to get correct and qualitative data into the DW. This process has basically the following tasks [10]:

978-1-4799-4233-6/14/$31.00 ©2014 IEEE

1) Converting data from heterogeneous data sources with various external representations into a common structure suitable for the DW.
2) Identifying and eliminating redundant or irrelevant data.
3) Transforming data to correct values
4) Reconciling differences among multiple data sources, due to the use of homonyms (same name for different things), synonyms (different names for same things).

B. Extraction, Cleansing, and Transformation Tools

1) Architectural Design and Modelling Design

At the beginning of ETL project, a decisive decision should be making about the architecture of the ETL system. Two choices are possible: (a) either to buy an ETL tool or (b) to build ETL processes from scratch. A survey conducted in [11] shows this disparity. Each policy has it pros and cons. While option (a) saves time, option (b) saves money. The intermediate option and solution will be open source tools like [7]. Designing ETL processes by the intermediate of an ETL tool or programming language is a technical task. The challenge is more in time and effort required to familiarize with the ETL tool in order for us to design the ETL process as states by Trujillo [12].

2) ETL Design Process

In [12], Trujillo et al propose a global process of ETL design. The authors propose a process composed on six actions: 1) select sources 2) transform the sources 3) Join the sources 4) select targets 5) map source to target attributes 6) Load data. Transformation step is a fundamental stage where this ETL process steps is further refined by Kimball [3, 4, 5] as cleaning and conforming steps. Without these two specifications, the end users will suffer from poor quality of data delivered by ETL [11, 16]. Hence, it is important that we can identify and correct “a large number of errors due to unanticipated values in the source data files” [11].

III. THE SURVEY-BASED SERVICES DATA WAREHOUSE SYSTEM DESIGN

This section describes the framework model for the Data Warehouse design shown in Figure 1.0. First, we obtain the business subjects on the type of analysis the reports can produced. Then we perform data quality check on data sources before we develop the data model and ETL for the data warehouse. Subsequently, we implement the ETL process in the development of the data warehouse. Lastly, we use the appropriate tool to perform the Business Intelligence analysis.

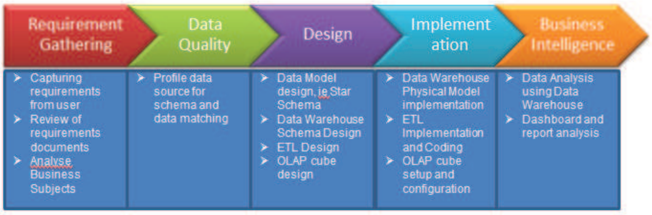

Requirement Gathering
Data Quality
Design
Implement ation
Business Intelligence
Capturing requirements from user
Review of requirements documents
Analyse Business Subjects
Profile data source for schema and data matching
Data Model design, ie Star Schema
Data Warehouse Schema Design
ETL Design
OLAP cube design
Data Warehouse Physical Model implementation
ETL Implementation and Coding
OLAP cube setup and configuration
Data Analysis using Data Warehouse
Dashboard and report analysis

Fig. 1. Data Warehouse Framework

A. Determination of analysis subject for the data model

Data in data warehouse was subject-oriented. Designing the data model for the data warehouse and analysis requires analyzing each subject area on the use of the data. In building of the data model, it is important that the data warehouse model excludes the organization’s daily operation data which is not used in the subject area for analysis. In each subject area section, we perform requirement gathering from the organization’s business users as well as screen shots of the system provided. Hence, we are able to determine the subject area for analysis by the respective business users. With the subject area determined, we can then build the respective data marts model for the data warehouse.

B. Data Quality

In the PoC (Proof of Concept), we perform quality checking in the area of the table schema matching the mapping documents provided as well as data checking in terms of Null Value, Range Value and Matching Value. The data profiling will help us to develop the respective data detection rules in order to determine the correct data to be inserted into the data warehouse for analysis. Poor quality of data [1,13] will affect the revenue of an organization. Typical data quality tasks include inconsistent of data references, missing fields [2] and empty value.

In this PoC, we are provided with the respective data source materials such as survey data source (MS Access MDB), mapping document, survey forms and screenshots of system. The data profiling methodology we conducted in this PoC are shown in the following steps:

1. Checking for Null Value in all columns of data
2. Checking for Inconsistency of data source value against the target tables.
3. Verifying the mapping document against between data source and target database.
    a. Analysis of the data source in Microsoft Access which is 300 columns in mapping document against 900 columns in the target tables.
4. Determining data relationship and assumption made by:
    a. Comparing of survey form with the tables and mapping document given.
    b. Determining linkage between the tables through the company registration number.
5. Creation of additional lookup tables such as gender, education level and industrial code definition.
6. Validating of Lookups:
    a. Industrial codes are specific to industry and it should match the survey form.
    b. State Codes

C. Data warehouse architecture design

The system was built according to the principle of three tier structure which is data acquisition tier, data storage tier and data display tier.

In data acquisition tier, data source are provided through MS Access MDB files. These data sources are extracted, cleansed (detection), transformed and finally loaded into the data warehouse[16]. In this PoC, we have developed data detection rules for the data exception handling.

In data storage layer tier, the physical representations of the data marts are developed. These data marts which are based on subject areas are developed with star schema and common dimensions in order to facilitate the cross-join drill down of the data used in the OLAP cube. The data stored in data marts have measures which are pre-aggregated. Further details of the data model design are given in Section IV.

In data display tier, data was analyzed and processed according to demands, and the analysis result can be shown by business intelligence and data mining tools. In this tier, accurate information that was required by business users can be made available for decision-making. Figure 2.0 illustrates the data warehouse architecture.

As shown in Figure 2.0, the data is stored in both the main data warehouse and data marts. The detailed data was stored in the data warehouse in the 3DNF. Data orientated to specific subject areas are stored in the respective data marts.

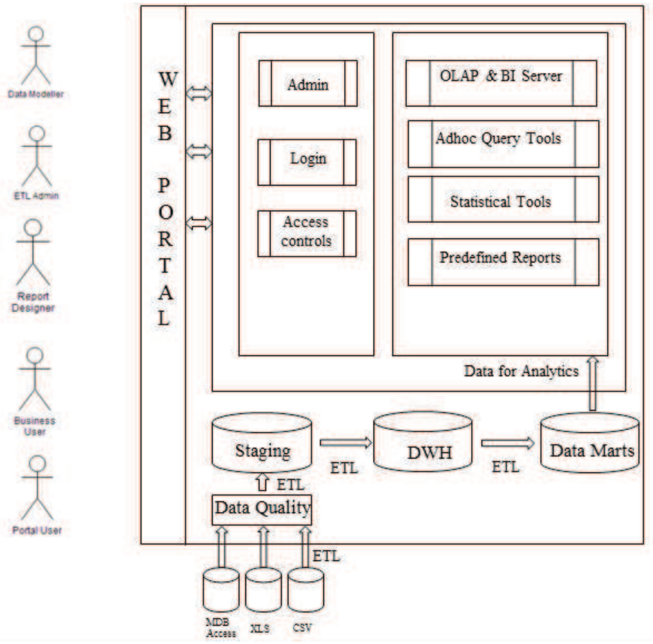


Fig. 2. Overview of Data Warehouse Architecture

IV. Data Model Design Consideration

A. Dimensional Modeling

Dimensional modeling is a logical design technique to organize dimensions and data marts in a star model. Every dimensional model is composed of a data mart table and a set of dimension tables (Bill Inmon/Ralph Kimball, 1997) [3, 4, 5, 6].

B. Dimension Tables

The dimension tables contain the textual descriptors of the business. In a well-designed dimensional model, dimension tables have many columns or attributes. These attributes describe the rows in the dimension table [3, 4, 5]. They are very useful in describing different entities in the business and provide textual meaningful data. They help to make business understandable.

C. Data Marts Tables

A data mart table is the primary table in a dimensional model where the numerical performance measurements of the business are stored [3, 4, 5]. It gives all the numerical measures at one place instead of being duplicated at different places in the data warehouse. They act as measures for analysis along various dimensions.

D. Attributes

Dimensional table contains collection of attributes which are useful for performing aggregations and analyzing business facts stored in fact table.

V. Star Schema Model for Survey-based Data Analytics

We have designed a multidimensional model using star schema for Survey-based Data Analytics shown in Figure 3.0 and Figure 4.0. The data marts are designed with ability to cater for more service based data in other areas besides the three services given in this PoC.

The data warehouse system have been implemented using a combination of tools such as Oracle 10G for the database, Pentaho Data Integration (or Kettle) for the ETL, DataCleaner Tool (For Data Profiling) and OLAP cube and Business Objects with MIMOS Mi-Bis.

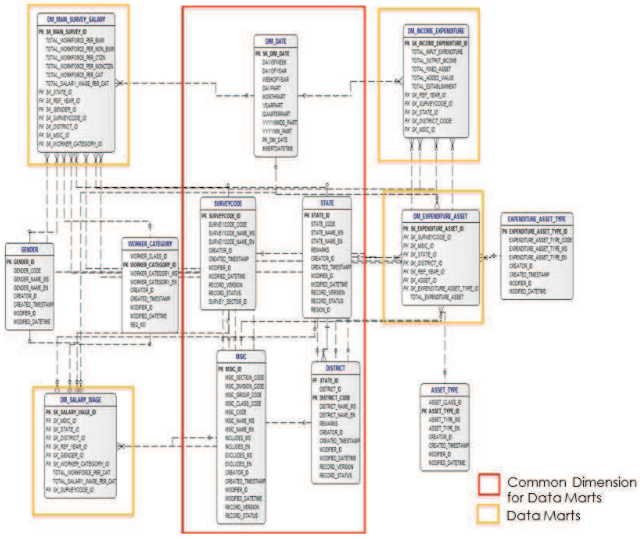


Fig. 3. Data Marts schema relationship model

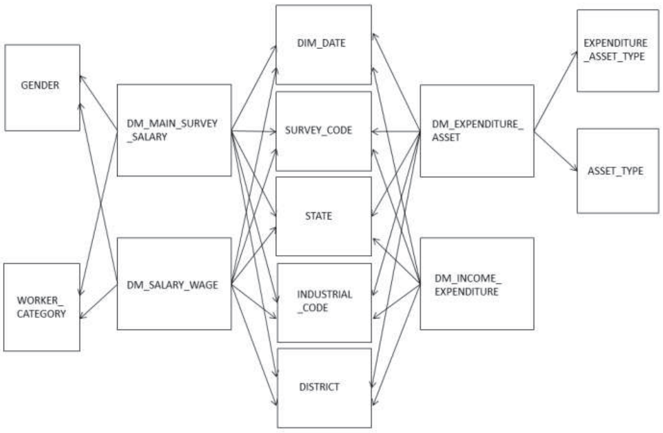


Fig. 4. Data Marts schema relationship model displayed in table name

The star schema consists of four data marts tables namely: Asset Expenditure, Salary Wage, Main Survey and Income. In general, the data marts are able to cater for ad-hoc reports on:

1. Number of workers and wage by worker category
2. Capital Expenditure by fixed asset
3. Detail of Statistical information by industry and states

Each data marts table is surrounded by dimensions for effectively analyzing the data. The following describes the common dimensions, dimensions surrounding the data marts and the data marts based on the star schema model in Figure 3.0 and Figure 4.0.

A. Common Dimensions

Dimension	Description
Date	It contains attributes describing various dates which are pre-generated with surrogate keys. The dates contains key attributes such as day of week, week of year, month, year, quarter, and others.
Survey Code	It contains attributes describing the survey code conducted on the services such as education, telecommunication, cargo and others.
Industrial Code	It contains attributes describing different type of industrial codes as per service sector.
State	It contains attributes describing each state where the industries are located. Hence drill down can be done according to state level.
District	It contains attributes describing each district where the industries are located.

B. Data Marts

Data Marts	Description
Expenditure and Asset	It contains attributes describing the expenditure and asset type with 1 measure and 7 different dimensions being linked.
Income	It contains attributes describing different type of income with 5 different measures and a linkages to 5 dimensions such as year, survey code, state, district and industrial code for different filtering options.
Main Survey	It contains attributes describing some of the workforce count based on citizen, non-citizen, ethnic, and worker category. It has 5 different measures and 7 different dimensions being linked.
Salary Wage	It contains attributes describing the salary distribution of the workforce with 2 measures and 8 different dimensions being linked.

C. *Dimension Tables Relating to Respective Data Marts*

Dimension	Description
Gender	It contains attributes describing the gender of the worker. It is used by Main Survey and Salary Wage Data Marts.
Worker Category	It contains attributes describing the worker category. It is used by Main Survey and Salary Wage Data Marts.
Asset Type	It contains attributes describing different type of asset type such as bus, lorry, computer software, etc. It is used by Expenditure and Asset Data Marts.
Expenditure Asset Type	It contains attributes describing different expenditure asset type such as new purchases, used purchased, gain loss and others. It is used by Expenditure and Asset Data Marts.

VI. ETL DESIGN PROCESS

The ETL processes are implemented using Pentaho Data Integration tool (or Kettle) [7] which uses Spoon GUI to help user define the ETL process to perform extraction and loading of the data source and transformed to the target database [15]. In the ETL processes, we also perform data profiling shown in Figure 6.0 which will be used to developed data detection rules for the data cleansing activities. The cleansed data is then subsequently transformed and loaded to the respective data marts shown in Figure 5.0 and Figure 7.0.

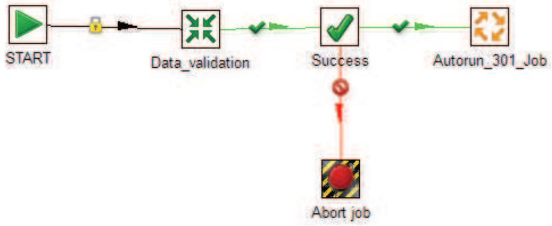


Fig. 5. ETL job from data loading to data warehouse with data detection rules.

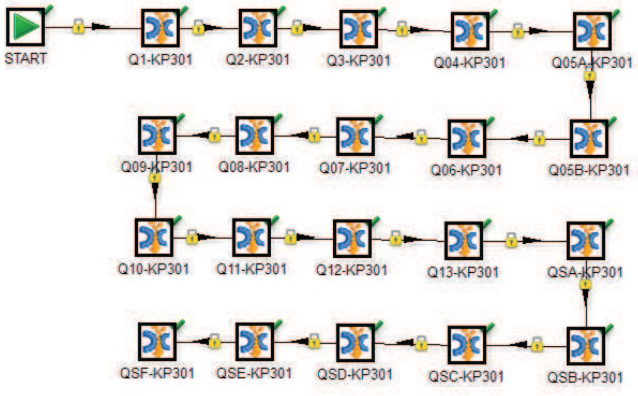


Fig. 6. ETL Job for data profiling

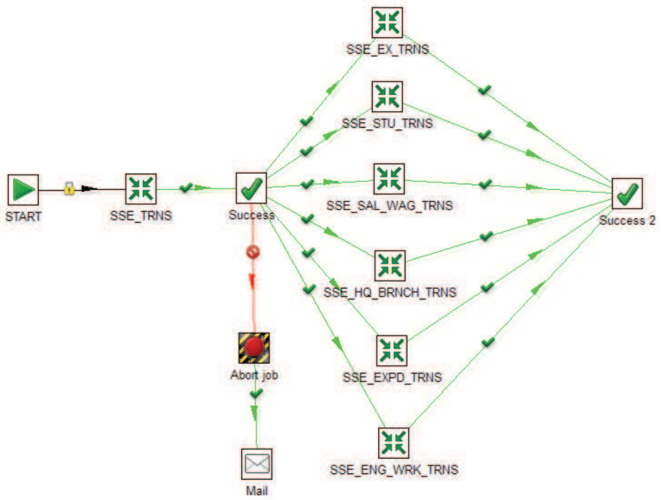


Fig. 7. ETL data transformation from source to target database

VII. RESULTS

We have successfully design and implemented the data warehouse using our data warehouse framework and the data model designed. In the data warehouse framework, we have developed specific data profiling methodology to address the challenges in this PoC, where the source data set are not provided with sufficient and correct information with respect to the relationship of the data for the reports. In the data model for the data warehouse, we have developed data model consisting of multiple data marts to support multi-dimension queries. The data model will be able to fulfill different queries based on the ad-hoc and pre-defined reports for the Survey-based data warehouse.

From Figure 3.0 and Figure 4.0, the data model has five common dimensions being shared by four data marts. The data model will enable the survey data from being analysed by data survey officer and management officials for their respective policy planning. Some of the key analyses that can be quarried onto the data models in the data warehouse are listed below:

- Total workforce by education

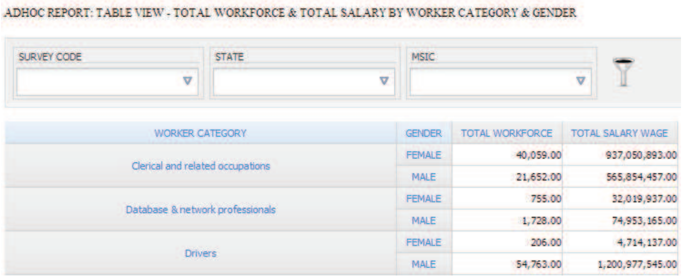
- Total workforce by telecommunication
- Total workforce by cargo services.
- Total workforce by gender based on education, telecommunication and cargo services.
- Year to year comparison of workforce growth based on education, telecommunication and cargo services.
- Top 10 industries with highest workforce employed in education service
- Top 10 industries with highest workforce employed in telecommunication service
- Top 10 industries with highest workforce employed in cargo service
- Total workforce by Worker Category based on education, telecommunication and cargo services
- Total workforce by Worker Category by gender based on education, telecommunication and cargo services
- Total workforce by Certificate Level such as PhD, Master, Degree, Diploma, etc based on education, telecommunication and cargo services
- Total workforce by Worker Category by Certificate Level
- Total Asset based on education, telecommunication and cargo services
- Top 10 industries with highest asset value in education service
- Top 10 industries with highest asset value in telecommunication service
- Top 10 industries with highest asset value in cargo service

In addition, the data model which uses the common dimension can enable multi-dimension queries to be made which can further broaden the scope of analysis.

Examples of multi-dimensional queries:

1. What is the total number of companies in more than one service?
2. What is the total workforce distributed by each service such as education, telecommunication and cargo services?
3. What are the top 10 industries irrespective of service sector having professional worker distributed by certificate level?
4. What is the total number of workforce having certificate level lower than diploma but hold a professional worker job?
5. What are the top 10 industries amongst the service sector having the highest paid salary?

The data warehouse system is a web-based dashboard system for displaying pre-defined reports, ad-hoc reports and data mining for analysis. The reports depicted in Figure 8.0, Figure 9.0 and Figure 10.0, display examples of the reporting data available in this PoC.



ADHOC REPORT: TABLE VIEW - TOTAL WORKFORCE & TOTAL SALARY BY WORKER CATEGORY & GENDER

SURVEY CODE:
STATE:
MSIC:

WORKER CATEGORY	GENDER	TOTAL WORKFORCE	TOTAL SALARY WAGE
Clerical and related occupations	FEMALE	40,059.00	937,050,893.00
	MALE	21,652.00	565,854,457.00
Database & network professionals	FEMALE	755.00	32,019,937.00
	MALE	1,728.00	74,953,165.00
Drivers	FEMALE	206.00	4,714,137.00
	MALE	54,763.00	1,200,977,545.00

Fig. 8. Ad-hoc Reports

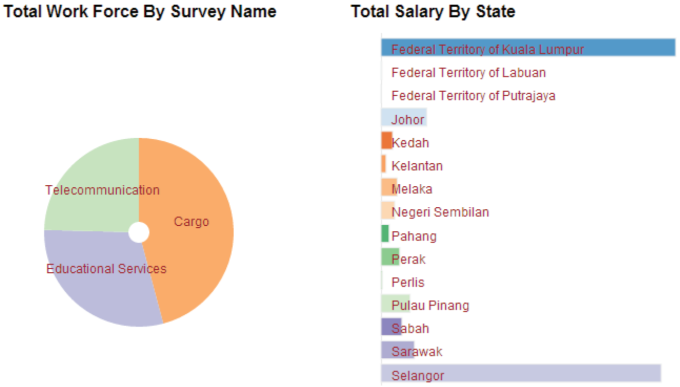


Fig. 9. Dashboard Charts Comparison

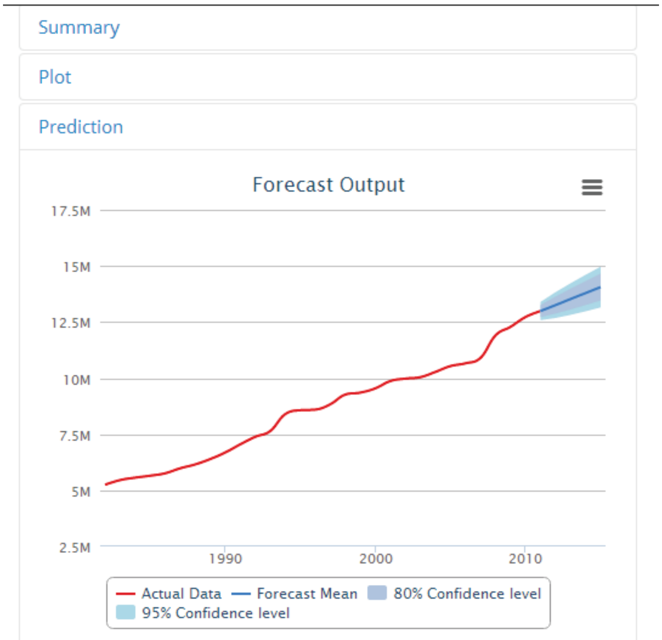


Fig. 10. Data mining report

VIII. CONCLUSION

In this paper, we demonstrated the design and implementation of data warehouse and data analytics using survey-based services data. The data warehouse is implemented using the data warehouse framework and data models we have designed as described in Section III, IV and V. The data models incorporate four data marts with multiple common dimensions to enable drill across the data marts for cross data analysis. Hence, multidimensional analysis can be performed such as employment based on education level, gender, state distribution and others amongst the education, cargo and telecommunication services. The authority can developed appropriate education and working policies to help address the needs of the industries. The data models are designed with ability to cater for more service based data in other areas besides the three services given in this PoC. In this data warehouse implementation, we also have designed the ETL process which incorporate the data quality to filter out inconsistent data with respect to data schema and data value. Hence, this data quality process will enable correct and qualitative data into the data warehouse.

REFERENCES

[1] R. Arora, P. Pahwa, S. Bansal, "Alliance Rules of Data Warehouse Cleansing", IEEE , International Conference on Signal Processing Systems, Singapore, May 2009, pp. 743 – 747.

[2] S. Chaudhuri, K. Ganjam, V. Ganti, “Data Cleaning in Microsoft SQL Server 2005”, In Proceedings of the ACM SIGMOD Conference, Baltimore, MD, 2005.

[3] R. Kimball, L. Reeves, M. Ross and W. Thornthwaite. " The Data Warehouse Lifecycle Toolkit : Expert Methods for Designing, Developing, and Deploying Data Warehouses", John Wiley & Sons, 1998.

[4] R. Kimball and M. Ross. "The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling", 2nd Edition, John Wiley & Sons, 2002.

[5] R. Kimball and J. Caserta. "The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data", John Wiley & Sons, 2004.

[6] W. H. Inmon, Building the data warehouse (2nd ed.), John Wiley & Sons, Inc., New York, NY, 1996.

[7] Pentaho Data Integration, www.pentaho.com/

[8] Apache Tomcat, http://tomcat.apache.org/

[9] T. Manjunath, S. Ravindra, and G. Ravikumar, “Analysis of data quality aspects in data warehouse systems,” International Journal of Computer Science and Information Technologies, Vol. 2, No. 1, 2010, pp. 477-485.

[10] B. Pinar, A Comparison of Data Warehouse Design Models, Master Thesis, Atilim University, Jan. 2005.

[11] W. Eckerson and C. White, “Evaluating ETL and Data Integration Platforms”, TDWI REPORT SERIES, 101communications LLC, 2003.

[12] J. Trujillo and S. Lujan-Mora. “A UML Based Approach for Modelling ETL Processes in Data Warehouses”. In I.-Y. Song, S. W. Liddle, T. W. Ling,and P. Scheuermann, editors, ER, volume 2813 of Lecture Notes in Computer science, Springer, 2003.

[13] Wang, Richard Y., and Strong, Diane M. Beyond accuracy: What data quality means to data consumers, Journal of Management Information Systems; Armonk; Spring 1996, 12 (4), pp. 5-34.

[14] Jarke, M., M. Lenzerini, Y. Vaasssiliou, P.Vassiliadis, "Fundamentals of Data Warehouse," 2000.

[15] You, Y. L. and Zhang, X. M., “A reliable strategy and design of architecture of ETL in data warehouse”, Computer Engineering and Applications, Vol. 10, 172_174 , 2005.

[16] B. K. Seah, "An application of a healthcare data warehouse system", Innovative Computing Technology (INTECH), 2013, pp. 269 - 273.