Student Performance Analytics using Data Warehouse in E-Governance System

S S Suresh
Asst. Professor, ASCT Department,
International Institute of Information Technology,
Pune, India

Mugdha Mahale
Student of School of Management,
International Institute of Information Technology,
Pune, India

ABSTRACT

Data warehouse (DWH) is a collection of databases consisting of cleaned and transformed data, integrated into subjective business areas for improved decision making. Many business organizations are enhancing their decision making capabilities using data warehouse as Decision Support System (DSS). For a company, data warehouse environment is very helpful in selecting right candidate at right time for the right job. For fresh college graduate/post graduate student's selection, integrated view across all aspects of students is essential. Data warehouse is the right choice to provide this integrated view with essential facts. Data warehouse environment effectively integrates information in an integrated repository, which can facilitate student-job mapping easier. Due to lack of sophisticated and effective DSS, many educational institutes are unable to showcase quality students to the industry. Human resource (HR) managers are finding difficulty in searching right college for procuring skillful students for the job. The current paper highlights the use of data warehouse with student information in E-Governance system. The current objective of this paper includes, designing data warehouse with multidimensional model within E-Governance architecture.

Keywords

Data warehouse, E-Governance, DSS, HR, multidimensional modeling

1. INTRODUCTION

In the current scenario businesses need real time information to make decisions, to stay ahead of the competition and to make efficient use of human and technological resources. Industry needs knowledge based systems to devise strategies, set goals, track and implement deviations in the normal course of actions during day-to-day operations. They can achieve this by implementing Decision Support Systems. A Decision Support System is a computer-based information system that supports business or organizational decision-making activities [1]. DSS serve the management, operations, and planning levels of an organization and help in making decisions. DSS facilitates the information in the organisation to have a single view i.e. be integrated. DSS enables the managers in the organisation to have accurate and precise information i.e. data integrity is preserved. The information in the organisation should have only one value throughout the organisation i.e. it gives it credibility. The information is available at any time thus timeliness of the information is ensured through the use of DSS. DSSs' are in use since 1970. There was much advancement that took place in the development of DSSs. In the middle and late 1980s, executive information systems (EIS), group decision support systems (GDSS), and organizational decision support systems

(ODSS) evolved from the single user and model-oriented DSS. [9]. In 1990s, data warehousing and on-line analytical processing began broadening the realm of DSS. For any DSS, database is a primary component. The underlying data model in conventional relational databases has limitations in analyzing large historical data effectively. This happens because the underlying data is normalized (E-R model, E. F. Codd). Hence, query processing and result finding consumes time. Also the E-R model is best suited for handling regular business operations. To effectively combat the problems of E-R model, data warehouse model was proposed by Bill Inmon (1990) (Father of Data Warehousing). In general, data warehouse supports multidimensional design (star schema) at the database level to query large pool of data. In this paper, we have given a thought to collecting and storing student data across all colleges (collectively) in a single integrated repository (data warehouse) and its use in E-Governance system. We have also proposed the architecture and it's working.

1.1 Data warehouse

Bill Inmon(1990) the father of data warehouse defined the data warehouse as: "a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's decision making process" [8]. The goal of the DWH is to establish a data repository that makes operational data accessible in a form that is readily acceptable for decision support and other user applications [3].The data warehouse is a consolidation of summarised information from heterogeneous sources. It contains aggregate and subject oriented data [8]. Data warehouse tools help support analysis and deriving information to various degrees of abstraction and granularity. The data warehouse manages the information and considers it to be a corporate asset. It contains voluminous, historical data and supports complex multidimensional queries. It has an enterprise wide scope and can be useful at departmental levels (data mart) as well. The data gets stored in the data warehouse at various levels (detailed and summarized/aggregated) which helps in quicker retrieval of data for decision making and optimizes query performance.

1.2 E-Governance

E-Governance implies the use of information communication technologies (ICT) by government bodies to impart information and deliver services to the net savvy citizens [7]. It is an effort to reach out to the public at a larger scale and bring transparency in the system. It enables the government to streamline its operations and allows the citizens to benefit and monitor the activities of the government and benefit from them. The concept of E-Governance exits in India through various initiatives like employment schemes, welfare schemes, education, online taxation, online forums and groups, development planning, agricultural produce analysis and

planning etc. It exists in various forms like Government to Citizen (G2C), Citizen to Government (C2G), and Government to government (G2G), Government to Business and Government to NGO (G2N) [6]. The focus of such an exercise is to shift paradigm from manual paper based systems to IT enabled efficient, consistent and effective decision making administrative systems. The E-Governance has three major focus areas:

- Increase efficiency in government processes

- Empower citizens

- Bring in transparency to the system, giving equal opportunities to the masses and reducing corruption.

1.3 HR System Problems

The HR personnel in a company are always on the lookout for talented and resourceful employees for their organisation. They need people who can be trained easily and who will be an asset to the company. The challenge the HR commonly face is finding the right people for the right jobs. Many companies visit colleges for campus placement. The HR personnel find it difficult to choose the right candidate for the suitable position. A concise overview of the skill set available across all colleges will help in making the recruitment process effective and efficient. In this paper we have proposed integration of student performance analytics data warehouse in E-Governance system. The proposed data warehouse will enable HR teams to access individual college data, skill sets and related student information. It will allow them to compare data from different colleges based on student performance, job experience, academic performance and skill sets of the students. Thus, this will facilitate easier screening and selection process. It also gives an integrated view about the all students as one single subject area.

1.4 Overview

The current paper deals with designing a multidimensional data warehouse model for student performance analytics. Section 2 deals with the proposed DWH Architecture. Section 3 focuses on design consideration. Section 4 deals with uses of the data warehouse from college, company and E-Governance point of view. Section 5 contains acknowledgements. Section 6 discusses conclusion followed by section 7 with references.

2. DWH ARCHITECTURE

The E-Governance architecture contains collection of applications and databases. There are various E-Governance architectures published (e.g. SOA E-Governance). In this paper, our aim is not to produce a new architecture. Instead, we are proposing data warehouse functionality to the any of the existing E-Governance architectures. We have proposed DWH containing college-student analytical facts and dimensions. The data for DWH would be extracted from various colleges'

operational data sources and then undergoes transformations. The transformed data would be integrated and loaded into data warehouse schema. The proposed DWH can be used by many different users like Government organization, placement officers in college, college directors/principles and HR teams from various organizations. The proposed architecture and its components are given below Figure.1

2.1 Working Scenario

2.1.1 Step1

The user/organization/company must register with the E-Governance system for accessing student data for recruitment. The registration module maintains user registration details. The system can be accessed online using web interface.

2.1.2 Step2

After completing the registration process, users can access services for which they have been registered. For example, if a user wishes to access student details for placement, he/she can access our proposed data warehouse using multidimensional query tools. The Access control module assigns priorities for accessing database tables. The data warehouse contains fact and dimensional tables for student performance analysis. The fact and dimensional tables are connected in a star schema model. The architecture maintains information delivery module as well. The information delivery module consists of OLAP (Online Analytical Processing), query and reporting tools. The OLAP tools allow easier multidimensional querying and reporting. It can be used as a self-reporting tool. The Query and Reporting tools allow users to fetch data from the DWH and aid in creating reports in various formats. Ad-hoc query tools support dynamic SQL constructed queries. For advanced analytics, Data Mining Query Language (DMQL) tools can be used. Generally, data mining algorithms are popularly used to search for patterns or trends stored in large databases [2].

2.2 Difficulties with Data Warehouse

The E-Governance difficulties in maintaining the data warehouse are given below:

- The architecture of this data warehouse is characterized by the size of the hardware platform, the number of transactional source systems and the number of users. This data warehouse will be connected to various separate operational systems; each of its users with different role might belong to a different organization, college, and government body.
- The ETL (Extract, Transform and Load) is a tedious and time consuming process. For any data warehouse development, ETL consumes 55%-60% of the total work. Hence, careful attention is required to design and schedule ETL jobs. Data checks, error conditions, integrity checks, null percentages and exceptional reports should be generated as per schedule.
- Developing and maintaining data warehouse is expensive as it requires many man-hours, time and resources.
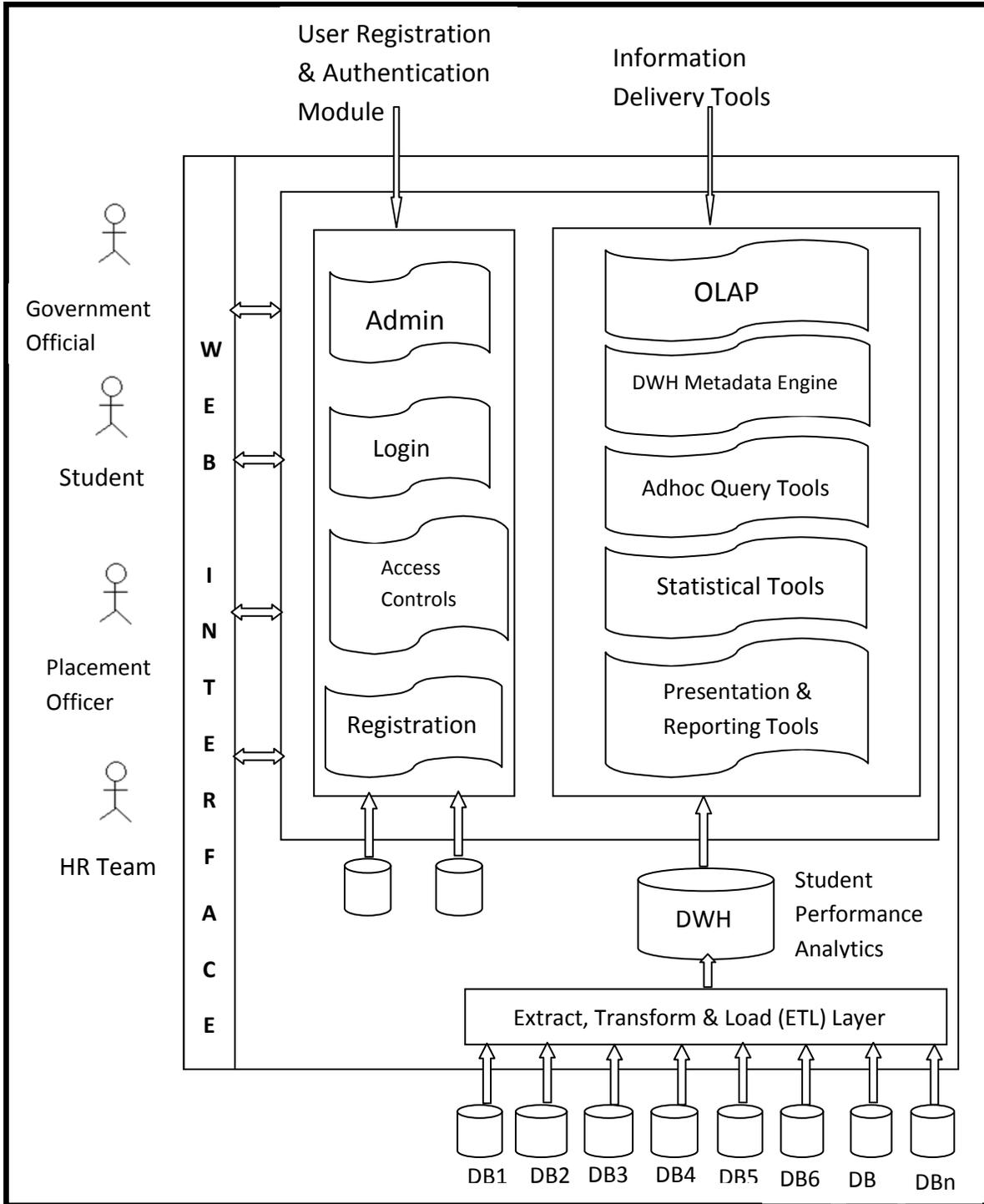
User Registration & Authentication Module

Information Delivery Tools

Government Official

Student

Placement Officer

HR Team

W
E
B

I
N
T
E
R
F
A
C
E

Admin

Login

Access Controls

Registration

OLAP

DWH Metadata Engine

Adhoc Query Tools

Statistical Tools

Presentation & Reporting Tools

DWH

Student Performance Analytics

Extract, Transform & Load (ETL) Layer

DB1   DB2   DB3   DB4   DB5   DB6   DB   DBn

Fig. 1: Architecture for E-Governance with Student Performance Analytics data warehouse

3. DESIGN CONSIDERATION

3.1 Dimensional Modeling

Dimensional modeling is a logical design technique to organize dimensions and facts in a star model. Every dimensional model is composed of a fact table and a set of dimension tables (Bill Inmon, 1997).

3.1.1 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 [5]. They are very useful in describing different entities in the business and provide textual meaningful data. They help to make business understandable.

3.1.2 Fact Tables

A fact table is the primary table in a dimensional model where the numerical performance measurements of the business are stored [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.

3.1.3 Attributes

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

3.2 Multidimensional Modeling

Data warehouse are based on multidimensional data model. Data is viewed in the form of a data cube. A data cube allows data to be modeled and viewed in multiple dimensions [2]. Using the OLAP facilities, users can perform multidimensional analysis and obtain multiple views of the data from multidimensional databases. This type of analysis is called slicing and dicing. Multidimensional views are inherently representative of any business model [4]. Multidimensional modeling also allows users to pivot the data, data aggregations, consolidation and classification across time dimensions. The multidimensional model thus has proved to provide high performance for queries and analysis.

3.3 Star Schema

Star Schema consists of (1) a large central (fact table) containing the bulk of the data, with no redundancy, and (2) a set of smaller attendant tables (dimension tables), one for each dimension [2]. The star schema consists of few fact tables referencing any number of dimension tables. Every dimension is equivalent; all dimensions are symmetrically equal entry points into the fact table. The logical model has no built-in bias regarding expected query patterns [5]. A star schema can be simple or complex. A simple star consists of one fact table and a complex star can have more than one fact tables. This type of schema is simple and symmetrical. This schema enhances and optimizes query performance [4].

3.4 Star Schema for Student Performance Analytics

We have designed a multidimensional model using star schema for Student Performance Analytics. It is yet to be implemented. We are planning to use SAS-Integration Studio 3.4 for materializing this schema. Implementation of this kind of project is possible only when colleges wish to share their data for integration with E-Governance system. Government intervention will be necessary to make data collection possible for such kind of projects so that the colleges share their data easily. The star schema consists of two fact tables namely: Student Performance Fact and Recruitment Fact. Each fact table is surrounded by dimensions for effectively analyzing the data. For the student analytics data warehouse, dimension tables and their description is given in Table 1 and schema diagram is given in Figure.2

Table 1. Dimensions

| Dimension	Description
Student	Contains attributes and its values describing students of various colleges
Academics	Contains attributes and its values describing academic details of college students
Course	Contains attributes and its values describing courses offered by various colleges and degrees awarded to the students
Job_Internship Details	Contains attributes and its values describing the work experience prior to joining the college and internship details of students completed during their course work
Academic Dates	Contains attributes and its values describing various academic dates in student academic history. This dimension facilitates searching students along time dimension
Job_Internship Dates	Contains attributes describing various dates related to student jobs and internships
HR Contact	Contains attributes and its values describing contact persons from different companies which have interacted with the placement team of the college who were involved in the recruitment process
Placement Dates	Contains attributes and its values describing student interview and placement dates
College	Contains attributes and its values describing college details
Student Skill set	Contains attributes and its values describing the key skills, students are marked on during interview process
Company Profile	Contains attributes and its values describing company details like job profile and company information
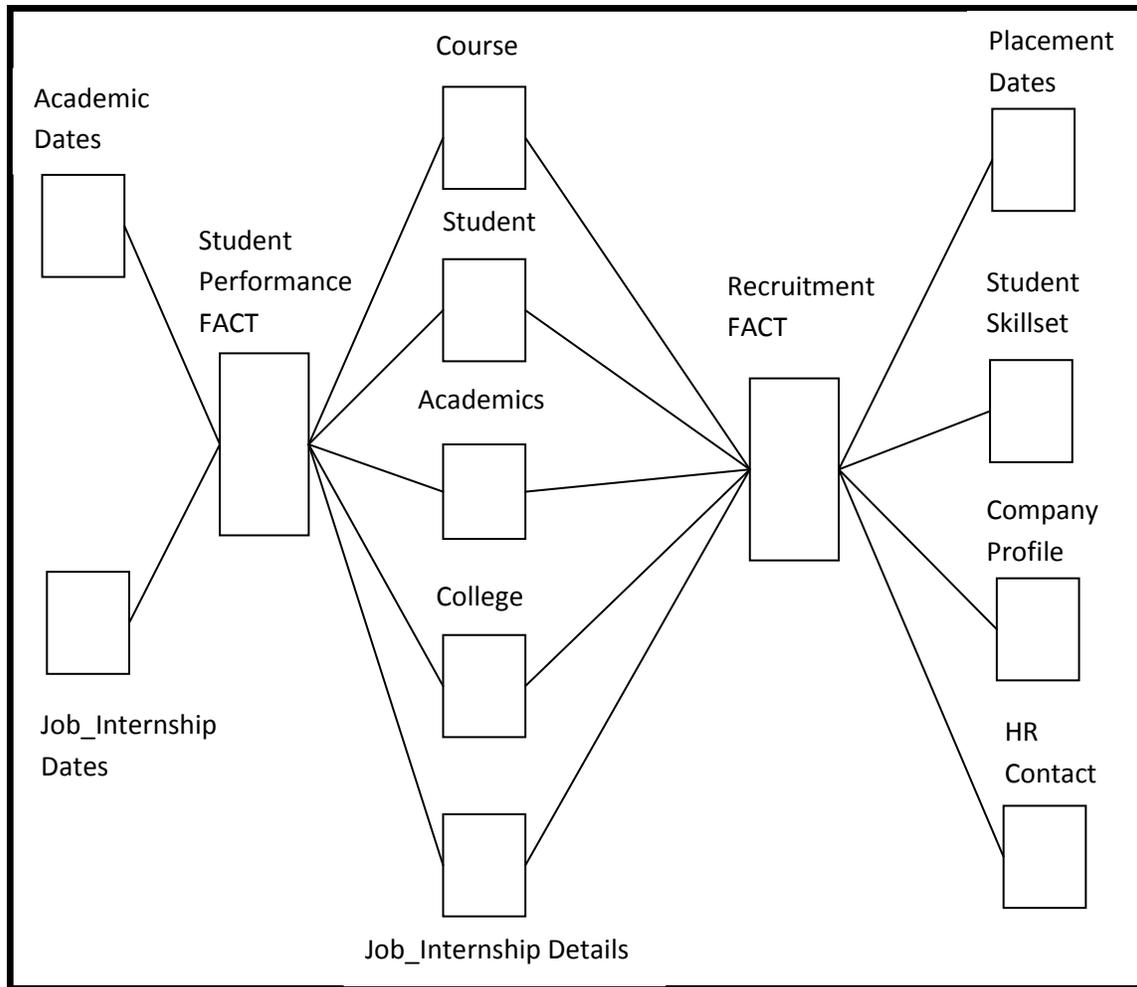
Fig. 2: Star Schema modeling for Student Performance Analytics Data warehouse.

From the Figure 2, student's performance can be gauged by HR, the placement officer, director/principal of any institute, Government officials etc. Course, Student, Academics, College and Job_Internship Details dimensions are conformed dimensions shared by the two fact tables. Let us Assume, student S1, enrolled in college C1, studying course CO1, academics key A1 and Job_Internship key JI1 for jobs and internships has AD1 and JID1 keys respectively for the Academic Dates and Job_Internship Dates tables respectively. Some of the vital facts which can be queried are listed below:

- Year1 Total marks

- Year2 Total marks

- Graduation Total marks

- Midterm examination Total marks

- Internal assessment Total marks

- Number of awards

- Number of certificates

- Number of scholarships

- Number of enrolled students

- Number of courses

- Number of degrees awarded

- Total scholarship amount

- Total attendance per semester

If the same student was placed in a company then we can get his Job/internship details from the HR/ placement team point of view through the recruitment fact. Say, he/she was placed in company COM1 for a certain job profile, SSK1 as the student skill set key, HR1 was involved in the recruitment process with placement officer PO1 from college and PD1 as the key describing his placement dates; The following are some of the measures that can be queried:

- Number of offers made to students

- Number of placed students

- Number of unplaced students

- Number of oncampusplaced students

- Number of offcampusplaced students

- Number of internship_unplaced students

- Number of internship_placed students

- Number of shortlisted students

- Number of resumes

- Number of students allotted to Placement officer

- Number of suitable candidates

- Number of unsuitable candidates

- Number of relevant skills

- Average salary package

- Max Salary package per institute

- Number of job roles

- Number of vacancies

- Number of filled positions

- Number of past_visits_campus

- Score in aptitude test

- Number of Successful_placements_campus

Thus we can track and add details for a student and find out his/her academic history and his/her placement history along a time dimension. We can track whether his academic skills and skill set requirements are matching the company or not. We can find out which student from a particular college was placed in which company for what profile and who were the HR and Placement Officers responsible for the process. What was the salary package offered? Which positions which are yet to be filled in a company? Which college to target based on student academics to get the right candidates for the right job? Which college to approach as the college conducts an industry oriented course? These are few of the questions which can be answered by the above schema.

Examples of Multidimensional queries:

1. What is the total number of students having graduated in the year 2009 with an engineering degree who have a prior job experience and are placed in a company X?

2. What is the total number of students and their scholarship amounts awarded, who are studying in the first year of masters program in networking and have scored above 3.5 GPA in college A?

3. What is the total number of students who have a prior internship experience, have excellent communication skills, completing post graduation in the year 2012, are enrolled in masters program in information technology and are placed for the profile of a Business Analyst in company X?

4. What is the total number of candidates from college X between the years 2008-2010, who are rejected because of poor analytical abilities?

5. Which colleges have maximum research papers and publications by students in the last year, who are yet unplaced?

4. USES OF DATA WAREHOUSE

The Student Performance Analytics data warehouse has many uses for different users in different roles;

4.1 College

- Students can view their academic history and take improvement decisions to perform better
- Director of the college can create a profile for the student
- He can measure the strengths and weaknesses of his students and get an overall picture of the performance
- He can gauge performance of students in a recruitment process for a certain company.
- He can monitor what courses are preferred more in the industry and make changes in the curriculum accordingly.

4.2 Company

- Tracking performance of the student at an individual level with regards to his academics and job/ internship details.
- Trying to find colleges and students who are marked high on their skill set matching points
- Trying to reduce the training cost
- Trying to reduce the probability of error in choosing right candidates.
- Background verification costs reduced as data available at college level itself.
- Since data warehouse maintains past historical data, companies can better understand the performance of college and its students.

4.3 E-Governance

- Maintaining skill set of young technical/ professional/skilled people in a country
- To track the number of students employed and unemployed after post graduation/ graduation.
- To monitor which is the most favoured fields and companies by students and colleges

5. ACKNOWLEGDEMENT

6. CONCLUSION

E-Governance is facilitating web and real time interaction between the government and the citizens smoothly. In this paper we tried to bring data warehouse environment to E-Governance. Using DWH the E-Governance operations can be performed at analytical level. Student dimension was used to understand different attributes like skills, marks and link them to measure the performance of a college. Recruitment process can be made more transparent, easier and reliable. With the help of this data warehouse the recruiter can choose skilled students and colleges at the right time. The Government can observe performance of colleges through various dimensions. The proposed E-Governance architecture facilitates better monitoring of

education system at various local areas. Despite having cost constraints, the concept has huge potential benefits in terms of multidimensional analysis and can help to solve India's industry need for skilled manpower.

7. REFERENCES

[1] 2011.Decision support system [online] (Last updated on: 31 January, 2011) available at http://en.wikipedia.org/wiki/Decision_support_system [Accessed on: 21st February, 2011]

[2] Jiawei Han and Micheline Kamber, Data mining concepts and techniques, 2nd edition, Elsevier Publications, 2005.

[3] Efraim Turban and Jay E. Aronson, Decision support systems and Intelligent Systems, 6th edition, PHI publications, 2000

[4] Paulraj Ponniah, Data warehousing fundamentals - A Comprehensive Guide for IT Professionals, John Wiley & Sons, Inc, 2001.

[5] Ralph Kimball, Data warehouse Toolkit, 2nd edition, Wiley & Sons, Inc, 2002

[6] Sachdeva Sameer (2003). E-governance Strategy in India. Management in Government, 34 (4): 35-53.

[7] Venkatesh S. (2003). Information and communication technologies: Impacts and impediments. Delhi: Authorspress. In Barthwal, C.P. (2003), E-governance for Good governance, The Indian Journal of Political Science, 64 (3-4): 285-308

[8] W. H. Inmon, "What is a Data Warehouse?" Prism Tech Topic, Vol. 1, No. 1, 1995

[9] Keen, P. G. W. (1978). Decision support systems: an organizational perspective. Reading, Mass., Addison-Wesley Pub. Co.