



## Warehousing and OLAP Analysis of Students Data-A Case Study

Saurabh Vijay\*, Saurabh Manek, Deepali Kamthania

Bhartia Vidyapeeth's Institute of Computer Applications and Management  
A-4 Paschim Vihar, Rohtak Road, New Delhi, India

---

**Abstract**— Educational data is incremental in nature. The large amount of data in educational institutions is generated in the form of personal information, academic data, placement data, fee payment data and much more. The problem is whenever there is a need to analyze or refer data for students' performance this turns out to be a cumbersome task as the data have been in different formats in different applications and some of the data is even maintained in hard copies over years. Educational institutions can apply data analytical techniques on the large amount of data that is generated within. A data warehouse is an important decision support system with cleaned and integrated data for knowledge discovery. In this paper an attempt has been made to design of data warehouse, the proposed system constitutes an integrated platform for a thorough analysis of student's past nine years data. The ETL process (extraction, cleaning, transforming and loading) have been performed with the help customized scripted tool. The tool helps in finding new relations and helps in creating reports. The results have been generated in the form of graphs which shows the trends and relationships. These trends and relations can be used to take decisions for improving student's performance and plan strategies and policies for progress and development of the Institute. Further the analysis of data has been achieved with online analytical processing OLAP operations.

**Keywords**— ETL (Extraction, transformation and loading), Data Warehouse, OLAP, MDX, Snowflake Schema

---

### I. INTRODUCTION

During recent years, universities have become more and more dependent on the collection, storage and processing of educational data. The huge amount of data stored in educational databases is increasing rapidly. The educational databases contain hidden useful information with many important factors related to the student's learning and performance. Since data obtained by any business processes commonly provide predictable clues about the future performances of systems that guide long-term investment plans for the assessment and restoration of process, this data should be safely stored while still being easily accessible for further analysis.

In the early 1970s, Morton, Sprague and Whinston developed a concept called decision support systems (DSS) [1, 2 and 3]. The birth of model-oriented DSS marked the beginning of information systems specifically for decision support in complex environments. In the mid 1980s, through integration with networking technology, artificial intelligence, and enterprise information systems, DSS such as distributed DSS, intelligent DSS, and integrated DSS appeared. In the 1980s, the concept of executive information systems (EIS) was developed. In the 1990s, the concept of on-line analytical processing (OLAP) systems was developed [4-8, 9 10, 11and 12]. In fact, in the 1990s, new technologies such as data warehouse, OLAP, and data mining consecutively emerged for DSS development; in which data warehouse concept emerged first. Following the introduction of data warehouses, OLAP and data mining appeared [13, 14 and 15]. In early 2000s, the most frequent research topics in data warehousing were the development of conceptual frameworks, designs, and system architectures. Delvin and Murphy from IBM [16] introduced the basic concept of the data warehouse to address various problems associated with business processes and information architecture that define the flow of data from operational systems to decision support environments. Originally, operational systems were developed to support daily business operations by maintaining and updating databases for order entry, billing, accounting and payroll. Delvin and Murphy [16] pointed out the limitations of operational systems for decision support and emphasized the need for more analytical information systems. Later, researchers modified this concept into the modern data warehouse. Inmon [17] established a concrete definition of a data warehouse, widely used in describing the basic features of data warehouses. Codd et al. [18] introduced the idea of OLAP to resolve problems that arise from the application of operational systems for decision support. OLAP is treated differently from online transaction processing (OLTP) with regard to the size, complexity, applicability, and time horizon of relevant data. Typically, OLTP focuses on day-to-day activities such as order entries and bank transactions which store specific values for individual fields [19]. On the other hand, OLAP handles values which represent a historical view of the entity over an extended time horizon [20]. Thus, as indicated by previous studies, a data warehouse provides a summarized and consolidated view of the relevant data rather than a detailed and individual view [19, 21].

The purpose of constructing a data warehouse is to provide a system that allows proper data to reach the right end user at just the right time [22]. To managers of an institution, the phrase “data warehousing” does not merely indicate an efficient tool for data integration but also implies the materialized format of a visualized, real-time management tool encompassing project design, rehabilitation and improvements. The multidimensional aspects of a data warehouse are well represented by a star schema and a fact table which defines all related dimensions. A fact table is located at the center of a star schema and includes two types of information dimension keys and facts [23]. In data warehousing, the multidimensionality of a star schema is often implemented by multidimensional cubes. These multidimensional features of the system allow for online analytical processing (OLAP) from a historical perspective. The objective of this study is to develop an in-built support system to provide a base for decision making within the Bharati Vidyapeeth’s Institute of Computer Application and Management (BVICAM) by analyzing nine years student’s data. A tool has been developed which can clean, transform, and load the unstructured data to a standard common format into a Data Warehouse for OLAP analysis. Further, a tool has been designed that assist in ETL process and load clean transformed nine year data for data ware so that OLAP analysis can be performed, the tool also generates report based on nine years student’s data. These reports can provide a base for decision making within the Bharati Vidyapeeth’s Institute of Computer Application and Management (BVICAM).

## II. STRUCTURAL DESIGN FOR WAREHOUSE

One of the critical processes in designing a data warehouse is to determine what data should be extracted from the various data sources to load into the data warehouse. Rujirayanyong and Shi [24] introduced two strategies for the identification of project data sources: the need-based approach and the availability-based approach. In the availability-based approach, any data about operational systems that is currently available is selected and uploaded to the data warehouse. The need-based approach, however, investigates the potential need for future analysis of data by considering the business nature of the system. In our study, the need-based approach was utilized to develop a data warehouse.

The next step is the structural design process. Usually, this process is composed of four steps, including selection of business processes, declaration of granularity, determination of associated dimensions, and identification of the facts [25]. The selection of business processes requires understanding of business requirements and associated data for the systems. Granularity refers to the required level of detail for information stored in a data warehouse [23]. In general, the level of detail is subject to the interests of users and to the amount of relevant data collected. Once the granularity of the fact table is determined, the related dimensions are reasonably obtained based on the nature of the fact table. Facts reflect the objective of the data warehouse and accordingly, should be composed of data useful for decision makers.

## III. MULTIDIMENSIONAL MODELLING FOR THE DATA WAREHOUSE

The multidimensional aspects of a data warehouse are well represented by a star schema and a fact table which defines all related dimensions. A fact table is located at the center of a star schema and includes two types of information — dimension keys and facts [23]. In data warehousing, the multidimensionality of a star schema is often implemented by multidimensional cubes. A typical dimensional model structure is a snowflake schema inherited from the star schema as shown in Fig.1 (a) and (b). The measurement of data can be recorded according to granularity. The data cube model provides a way to aggregate facts along multiple attributes called dimensions. In the data cube, data is stored as facts and dimensions instead of rows and columns as in a relational data model. Table is the data model of data warehouse. In the dimension model, all tables are concluded in two types: dimensional table and fact table. The fact table for student performance provides the Sub\_avg, sub\_distinction\_no, sub\_1<sup>st</sup>\_division\_no, sub\_2<sup>nd</sup>\_division\_no, sub\_below\_50\_no of students, the related three dimensions are batch and subject and company. The fact table for placement provides the eligible\_students, appeared\_students, placed\_students. The related three dimensions are batch, subject and company.

### A. Snowflake Schema

In the current case the Batch dimension of the BVICAM’s warehouse has been normalized to form a batch\_performance dimension table which can be seen in the following diagrams (Fig.1).

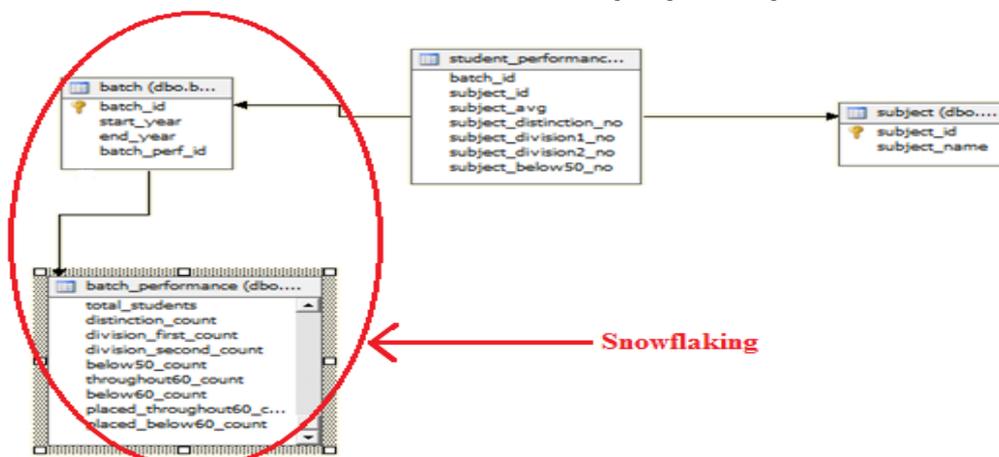


Figure 1 (a): Snowflake Schema of Student Performance

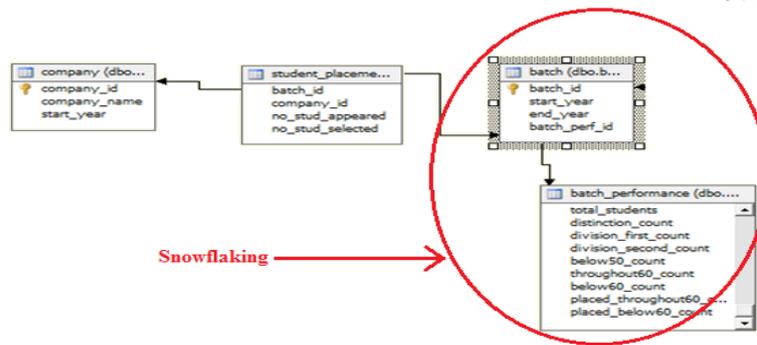


Figure 1 (b): Snowflake Schema of Student Placement

### B. Family of Stars

In the current context of BVICAM's data, two star schemas have been presented. The following family of stars has been obtained by joining the schemas (refer Fig. 2).

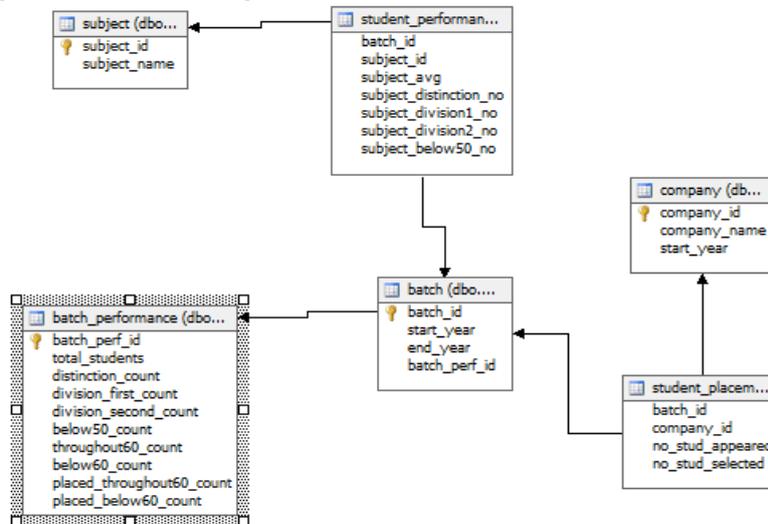


Figure 2: Family of Stars

## IV. DATA EXTRACTION TRANSFORMATION AND LOADING IN DATA WAREHOUSE

Information integration is one of the main problems to be addressed when designing a Data Warehouse. Possible inconsistencies and redundancies between data residing at the operational data sources and migrating to the Data Warehouse need to be resolved, so that the warehouse is able to provide an integrated and reconciled view of data within the organization. The basic components of a data integration system are wrappers and mediators. A wrapper is a software module that accesses a data source, extracts the relevant data, and presents such data in a specified format, typically as a set of relational tables. A mediator collects, cleans, and combines data produced by wrappers and/or other mediators, according to a specific information need of the integration system. The specific and the realization of mediators is the core problem in the design of an integration system. The data stored in the Data Warehouse should reflect such an informational need, and hence should be defined in terms of the corporate data.

### A. ETL Process (Extract-Transform-Load)

ETL comprises a process of how the data are loaded from different source systems to the data warehouse. Currently, the ETL encompasses a cleaning step as a separate step [26]. The sequence is then Extract-Clean-Transform-Load.

#### IV.A.1 Extraction

In the present study the data of student performance and placement was present in various different formats of MS excel, word files and hard copies. This raw data need to be compiled to a common format in an MS excel file. The data has been changed to common standard format that is to be used for the further cleaning and loading process. The standard format of performance of students for a particular batch according to the start year of that batch is shown in Fig. 3 (a).

- Column A: roll number of MCA.
- Column B: name of student.
- Column C: class 10<sup>th</sup> aggregate percentage.
- Column D: class 12<sup>th</sup> aggregate percentage.
- Column E: name of company in which student got placed.
- Column F- [...]: marks secured in particular subject ID.

The standard format of placement of students for a particular batch according to the start year of that batch is shown in Fig 3 (b).

- Column A: name of company.
- Column B: number of students appeared in that company.
- Column C: number of students placed in that company.

	A	B	C	D	E	F	G	H	I	J
1	Roll	Name	10th	12th	Grad	placed	S101	S103	S105	S107
2	02/BVMC/2002	MEENAKSHI GARG	76	74	73	HCL	80	79	80	57

Figure 3 (a): Standard Excel Format of Student Performance

	A	B	C
1	company name	Appeared	Selected
2	HCL	30	9
3	DCM	13	1

Figure 3 (b): Standard Excel Format of Student Placement

#### IV.A.2 Cleansing

The data cleansing was required as the provided raw sheets had some of the anomalies like missing values, or the inconsistent format of values, or null values inserted. In order to clean the extracted data present in the standard excel format, the missing values were removed by using average out values for that places. The null values were replaced by the value '0' in the excel file.

#### IV.A.3 Transformation and Loading via (Scripted Tool)

The transformation and loading of data in the target data warehouse i.e. database table present in the SQL Server 2005 has been done through the scripted tool. In this case transformation was required as the extracted excel sheets for each batches had same format but the meaning was different. Since the MCA syllabus has changed in various years and so the meaning of several subject ID's. The following figure shows the change in subject code in different batches. The transformation has been done with the help of mapping algorithm developed in order to transform the data of 2002 syllabus, 2004 syllabus and 2010 syllabus to a common format having the same subject ID for each batch. A Dictionary has been used to map the subject ID with their names. The mapping process is shown in fig 4(c).

Fig. 4 (a) shows the subject id S101 in 2002 is 'programming in C' whereas in 2004 syllabus it is 'Fundamentals of IT' as shown in Fig. 4 (b). From this mapping process Fig. 4 (c) shows that subject ID (S101) in 2002 and subject ID (S105) in 2004 is made to subject ID (S103) in syllabus for all the batches. The mapping has been done with the script before loading the data into the database tables.

```

("S101", "Programming in C");
("S103", "Digital Electronics");
("S105", "Discrete Mathematics");
("S107", "Organizational Behaviour");
("S109", "Financial Accounting");
("S151", "Practical - I");

```

Figure 4 (a): Syllabus of year 2002

```

("S101", "Fundamentals of IT");
("S103", "Digital Electronics");
("S105", "Programming in C");
("S107", "Discrete Mathematics");
("S109", "Financial Accounting");
("S151", "Practical - I");
("S153", "General Proficiency - I");

```

Figure 4(b): Syllabus of year 2004

```

("Fundamentals of IT", "S101");
("Programming in C", "S103");
("Discrete Mathematics", "S105");
("Computer Organization", "S107");
("Principles and Practices of Management", "S109");
("Fundamentals of IT Lab", "S151");
("Programming in C Lab", "S153");
("Computer Organization Lab", "S155");
("General Proficiency - I", "S161");

```

Figure 4 (c): Syllabus for common format

The tool provides an interface as shown in Fig. 5 which can transform and load the raw data into SQL Server 2005 database tables present in the specified format. The steps of loading are:

Step 1: Choose the add BVICAM data option from the available options. Click OK button.

Step 2: Enter the information of particular batch and chose the excel file and click on load and clean data button in order to transform and load the data into database.

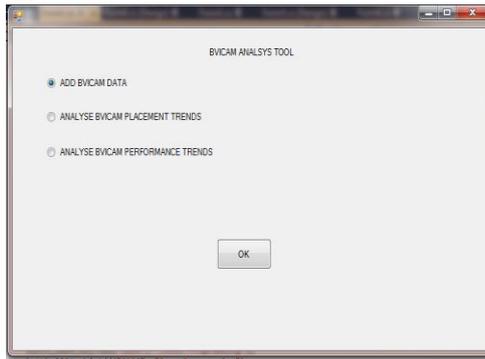


Figure 5(a): BVICAM Analysis Tool Interface

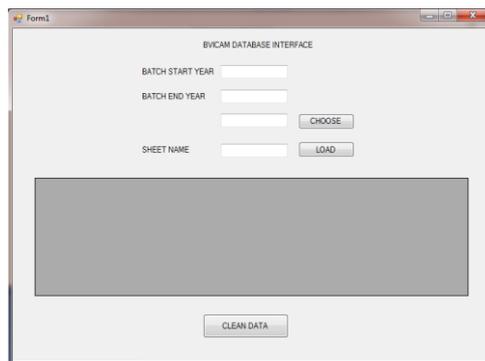


Figure 5(b): BVICAM Data Loading Interface

## V. OLAP (ONLINE ANALYTICAL PROCESSING)

**V. OLAP** is a way of answering multi-dimensional analytical queries. The OLAP encompasses the traditional relational system and provides a multi-dimensional view of the data for analysis. OLAP has a multi-dimensional cube or hypercube at the heart. The OLAP cube consists of facts and measures, which are categorized by dimensions.

### V.A.1 OLAP Cube Creation

OLAP cubes are hypercube that incorporate data in more than 2 dimensions. OLAP cubes allow using multiple dimensions. The query language used for interacting with OLAP cubes is MDX (Multi-dimensional expressions). In the current case of BVICAM data, the cube creation is done based on the snowflake schema designed earlier. The data from the SQL Server 2005 has been imported into the Business Intelligence 2005 which is used for cube creation. Fig. 6 shows data cube for BVICAM Warehouse

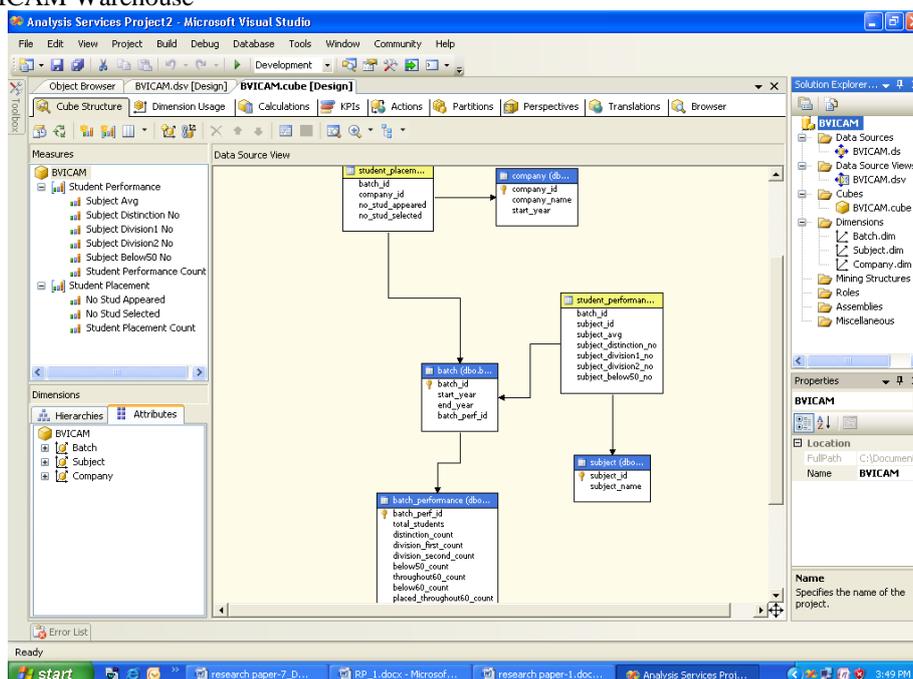


Figure 6: Data Cube for BVICAM Warehouse

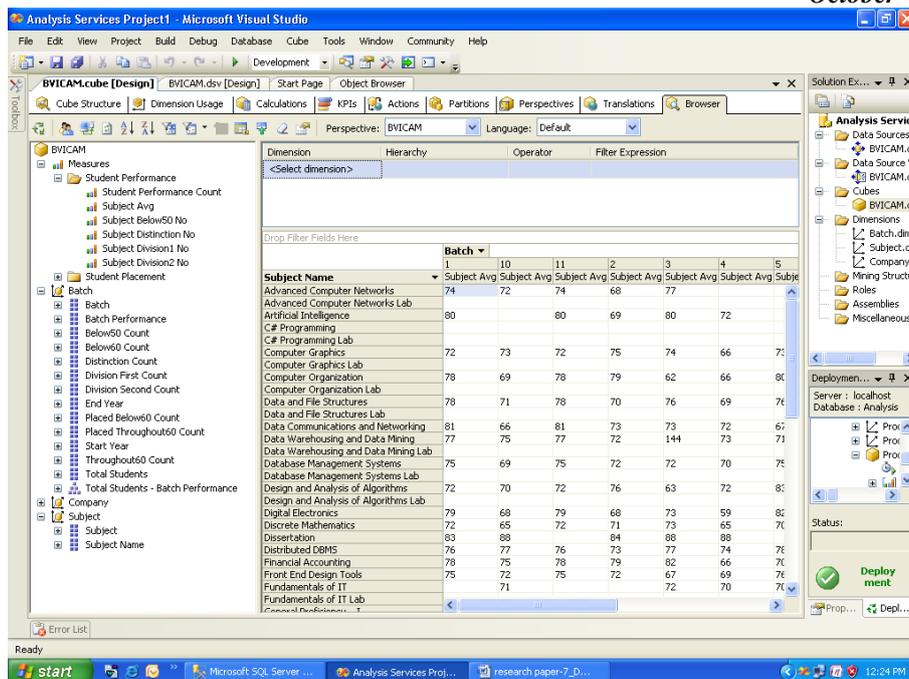


Figure 7: Average of different subject for 10 batches.

Fig. 7 shows average marks of subject for nine batches of BVICAM. Similarly other data can be viewed from different dimensions

#### V.A.2 OLAP using MDX (Multi-Dimensional Expressions)

MDX is a query language for OLAP databases as SQL is for relational databases. There are six primary data types in

- MDX: Scalar- Neither number or a string, can be defined as a literal.
- Dimension/Hierarchy- The dimension or the hierarchy of dimension of the cube.
- Level- Level in the dimension hierarchy.
- Member- Member in a dimension hierarchy.
- Tuple- An ordered collection of one or more members from different dimensions.
- Set- An ordered collection of tuples with the same dimensionality.

Use of some of the MDX queries has been done in the current project to generate the results of some OLAP operation on the data. The MDX queries and their results are shown in the Result and Discussion Section.

### VI. RESULTS OF MDX QUERIES

1. Select [Subject].[Subject Name].Members on 0,[Batch].[Start Year].Members on 1From [BVICAM] where {[Measures].[Subject Avg]}

The result of the query is shown in Fig. 8 which shows the average marks of students or the average performance of the students in each subject for every year. This can help us to know the where the students actually lacked.

	All	Advanced Computer Networks	Advanced Computer Networks Lab	Artificial Intelligence
All	24842	438	86	301
2002	2532	74	(null)	80
2003	2454	68	(null)	69
2004	3095	77	(null)	80
2005	2846	(null)	(null)	72
2006	2510	(null)	(null)	(null)
2007	2757	(null)	(null)	(null)
2008	2785	74	(null)	(null)
2009	2818	72	(null)	(null)
2010	3045	73	86	(null)

Figure 8: Average Marks of Students in Various Subjects Every Year

2. Select [Company].[Company Name].Members on 0,[Batch].[Start Year].Members on 1From [BVICAM] where {[Measures].[No Stud Selected]}

The above query is used to calculate the total number of students placed till now in each of the company that has ever been recruiting the students of this organization. This information helps to know the recruitment status of the students and in which companies are the student most placed till now as shown in Fig. 9.

	All	3pillar global	absolute data	accenture	alcatel	aon hewit	aricent	aspiring minds
All	451	1	8	18	2	2	69	0
2002	29	(null)	(null)	(null)	(null)	(null)	(null)	(null)
2003	41	(null)	(null)	(null)	(null)	(null)	(null)	(null)
2004	46	(null)	(null)	(null)	(null)	(null)	(null)	(null)
2005	88	(null)	3	18	(null)	(null)	12	(null)
2006	45	(null)	(null)	(null)	2	(null)	20	(null)
2007	39	(null)	0	(null)	(null)	(null)	10	(null)
2008	61	(null)	5	(null)	0	(null)	(null)	(null)
2009	67	(null)	(null)	(null)	(null)	(null)	23	0
2010	35	1	(null)	(null)	(null)	2	4	(null)

Figure 9: No. of Students Selected in Various Companies Every Year

3. Select {[Measures].[No Stud Appeared],[Measures].[No Stud Selected]} on 0, [Batch].[Start Year].Members on 1From [BVICAM]

The above query calculates the total number of opportunities for every batch and the number of successful placements out of those as shown in Fig. 10. It is also helpful to calculate the ratio of selection against the opportunities made available to the students.

	No Stud Appeared	No Stud Selected
All	3874	451
2002	156	29
2003	455	41
2004	176	46
2005	475	88
2006	439	45
2007	484	39
2008	670	61
2009	464	67
2010	555	35

Figure 10: Total No. of Opportunities And Students Placed Every Year

4. Select [Subject].[Subject Name].Members on 0,[Batch].[Start Year].Members on 1From [BVICAM] where {[Measures].[Subject Distinction No]}

	All	Advanced Computer Networks	Advanced Computer Networks Lab	Artificial Intelligence
All	10198	92	55	43
2002	657	15	(null)	23
2003	1107	11	(null)	9
2004	1146	27	(null)	6
2005	1154	(null)	(null)	5
2006	1080	(null)	(null)	(null)
2007	1122	(null)	(null)	(null)
2008	1168	10	(null)	(null)
2009	1231	5	(null)	(null)
2010	1533	24	55	(null)

Figure 11: Distinctions of Students in Various Subjects Every Year

This query shows the total number of distinctions scored in the every subject throughout the years. This helps to know the subjects which get the highest distinctions as shown in Fig. 11. This will help in segregating the subjects which are easy to score for most of the students of the batch. Student with low percentage can work on to score in these to increase the overall performance.

## VII. CONCLUSIONS

The prototype model that has been developed for analysis on nine years BVICAM student performance and placement data can be extended to different institutions across the university running same course for the analysis. The execution of MDX queries on the BVICAM academic and placement data generated some interesting results. It has been observed that the mathematical, calculative and algorithm based subjects are low scoring as compared to other subjects. The subjects like practical's and dissertation prove to be the subjects where almost all students score distinction. The placement record clearly shows that organizations like TCS, Aricent hire more students almost every year. The institute should keep in mind of healthy relationship with these companies for the future as this can be beneficial at the time of Industry Institute Partnership Collaboration. Several companies have lower hiring ratio from BVICAM because of the inconsistency of the visits. Another important thing to notice is that the companies which come in the campus for recruitment hire more than those who call the students elsewhere for off campus recruitment or pool campus drives.

REFERENCES

- [1] M.S. Gorry, Scott Morton, "A framework for management information systems", Sloan Management Review, Vol. 13 No.1 pp 50–70 (1971).
- [2] C.W. Holsapple, A.B. Whinston, "A decision support system for area-wide water quality planning", Socio-Economic Planning Sciences, Vol. 10 No. 6, pp 265– 273 (1976).
- [3] J. Tian, Y.L. Wang, H.Z. Li, L.X. Li, K.L. Wang, DSS development and applications in China, Decision Support Systems doi:10.1016/j.dss.2004.11.009.
- [4] S. Miksch, Artificial intelligence for decision support: needs, possibilities, and limitations in ICU", in: A. Gullo (Ed.), Anaesthesia, Pain, Intensive Care, and Emergency. Medicine (APICE-95), Proceedings of the 10<sup>th</sup> Postgraduate Course in Critical Care Medicine, Springer, Berlin, pp. 901–908 (1995).
- [5] R.H. Mohring, R. Muller, F.J. Radermacher, Advanced DSS For Scheduling: Software Engineering Aspects and the Role of Eigenmodels, Proceedings of the 27<sup>th</sup> Annual Hawaii International Conference on System Sciences, Maui, HI, (1994).
- [6] A. Needamangala, A library decision support system built on data warehousing and data mining concepts and techniques". Thesis for Master's Degree, University of Florida, (2000).
- [7] J.L. Pollock, "OSCAR-DSS", OSCAR Project Technical Report, (1997).
- [8] M. Postema, T. Menzies, X. Wu, Decision support tool for tuning parameters in a machine learning algorithm", PACES/.SPICIS '97 Proceedings, Nanyang Technological University, Singapore, pp. 227– 235 (1997).
- [9] Z. Shi, Y. Huang, Q. He, L. Xu, S. Liu, L. Qin, Z. Jia , J. Li, H. Huang, L. Zhao, MSMiner—a developing platform for OLAP, Decision Support Systems 42 2016–2028 (2007).
- [10] F. Zelezny, J. Zidek, O. Stepankova, "A learning system for decision support in telecommunications", Proceedings of the 1<sup>st</sup> International Conference on Computing in an Imperfect World, Belfast (2002).
- [11] Y. Zhu, C. Bornhvd, D. Sautner, A. Buchmann, "Materializing web data for OLAP and DSS, 1<sup>st</sup> International Conference on Web-Age Information Management", WAIM'00, Shanghai, China (2000).
- [12] E. Blanzieri, P. Giorgini, P. Massa, S. Recla, "Data mining, decision support and meta-learning: towards an implicit culture architecture for KDD", Proceedings of the Workshop on Positions, Developments and Future Directions in Connection with IDDM-2001 (2001).
- [13] H.X. Li, L. Xu, "Feature space theory—A mathematical foundation for data mining", Knowledge-Based Systems Vol. 14 253– 258 (2001).
- [14] H.X. Li, L. Xu, J. Wang, Z. Mo, "Feature space theory in data mining", Expert Systems 20, 60– 71 (2003).
- [15] P. Shim, M. Warkentin, J.F. Courtney, D.J. Power, R. Sharda, C. Carlsson, Past, present, and future of decision support technology, Decision Support Systems 33, 111 –126 (2002).
- [16] B.A. Devlin, P.T. Murphy, An architecture for a business and information system, IBM Systems Journal 27 (1) 60–81 (1988).
- [17] W.H. Inmon, Building the Data Warehouse, John Wiley & Sons, Inc., New York, (1992).
- [18] E.F. Codd, S.B. Codd, C.T. Sally, Providing On-line Analytical Processing to User-analysis: an IT mandate, E.F. Codd and Associates, San Jose, California, (1993).
- [19] S. Chaudhuri, U. Dayal, An overview of data warehousing and OLAP technology, ACM SIGMOD Record 26 (1) (1997).
- [20] S. Conn, OLTP and OLAP data integration: a review of feasible implementation methods and architectures for real time data analysis, in: The Proceedings of IEEE Southeast, April 8–10 (2005).
- [21] D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi, R. Rosati, Data integration in data warehousing, International Journal of Cooperative Information Systems 10 (3) (2001) 237–271.
- [22] K.W. Chau, Y. Cao, M. Anson, J.P. Zhang, "Application of data warehouse and decision support system in construction management", Automation in Construction Vol. 12 No.2 (2002), 213–224.
- [23] R.J. Roiger, M.W. Geatz, "Data Mining: A Tutorial-based Primer", Pearson Education, (2003).
- [24] T. Rujirayanyong, J.J. Shi, A project-oriented data warehouse for construction, Automation in Construction 15 (6) 800–807 (2006).
- [25] R. Kimball, M. Ross, The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, 2nd ed. John Wiley & Sons, Inc., New York, 2002.
- [26] <http://homepages.inf.ed.ac.uk/wenfei/tdd/reading/cleaning.pdf>