IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 5, May 2017

Data Warehousing Architecture and Pre-Processing

Vishesh S[1], Manu Srinath[1], Akshatha C Kumar[2], Nandan A.S.[3]

B.E, Department of Telecommunication Engineering, BNM Institute of Technology, Bangalore, India [1]

Student, Department of Telecommunication Engineering, BNM Institute of Technology, Bangalore, India [2]

Student, Department of Computer Science and Engineering, BNM Institute of Technology, Bangalore, India [3]

Abstract: The computerization of our society has substantially enhanced our capabilities for both generating and collecting data from diverse sources. A tremendous amount of data has flooded almost every aspect of our lives. There is a need in transforming the vast amount of data into useful information and knowledge. This has led to the generation of promising and flourishing frontier in computer science called data mining. Data mining is the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the web, other massive information repositories or data streams. Data mining can be applied to any kind of data as long as the data is meaningful for a target application. In this paper, we discuss in detail data warehouse and data warehouse data, which is almost basic form of data for data mining applications. We also present to you a typical framework of a data warehouse and data pre-processing techniques. We also discuss about OLAP (Online Analytical Processing) Data Marts which is a subset of an organizational data store, usually oriented to a specific purpose or major data subject, which may be distributed to support business needs.

Keywords: computerization, data mining, databases, data warehouses, data pre-processing techniques, OLAP (Online Analytical Processing) Data Marts.

I. INTRODUCTION

Database and information technology has evolved comprehensively from primitive file processing systems. Advanced database systems, data warehousing and data mining [1] for advanced data analysis and web-based databases [2] incorporate new and powerful data models. The steady and dazzling process of computer hardware technology and powerful processing, affordable data collection equipment and storage media has catalysed data mining and analytics. One of the emerging data repository architecture is the data warehouse. This is a repository of multiple heterogeneous data sources organized under a unified schema at a single site to facilitate management decision making. Figure 1 shows the typical framework of a data warehouse. Suppose Konigtronics Private Limited is a successful international company with many branches around the world and each branch has its own set of databases. The relevant data is spread over several databases physically located at numerous sites. To provide an analysis, consider company sales of a particular quarter of a financial year. It would be a difficult task to analyse the financial data spread over different databases. Having a data warehouse would dilute the problem statement. A data warehouse is a repository of information collected from multiple sources (branches) and unified in a single site as shown in the figure 1. Suppose Konigtronics Private Limited has branches in several places like Bangalore, New York, London and Spain, all the data from these sources should be pre-processed before unifying it into the data warehouse. Following are the data processing techniques:

- Data cleaning
- Data integration
- Data reduction
- Data transformation
- Load and refresh

Data cleaning can be applied to remove noise and correct inconsistencies in data. Integration merges data from multiple sources into a coherent data store such as a data warehouse. Data reduction can reduce data size by aggregating, eliminating redundancies or clustering. Data transformations are a normalisation process to scale the data to fall between 0.0 and 1.0. Specific query and analysis tools are used in mining data from data warehouse and projected to clients. OLAP is an acronym for Online Analytical Processing. OLAP is used to perform multidimensional analysis of business data and provides capability for complex calculation, trend analysis and sophisticated data modelling. It is the foundation for many kinds of business applications for business performance management and marketing. Planning,

DOI10.17148/IJARCCE.2017.6503

budgeting, forecasting, financial reporting, analysis, simulation models, knowledge discovery and data warehouse reporting.

II. DATA WAREHOUSE

Data warehouses generalise and consolidate data in multi-dimensional spaces. The construction of data warehouses involves data cleaning [3], data integration [4] and data transformation [5] and the data can be viewed as an important step for data mining. Data warehouse provides architectures and tools for business executives to systematically organize, understand and use their data to make strategic decisions. The information gathered in a warehouse can be used in any of the following domains:

- Tuning production strategies
- Customer analysis/customer behaviour
- Operations analysis

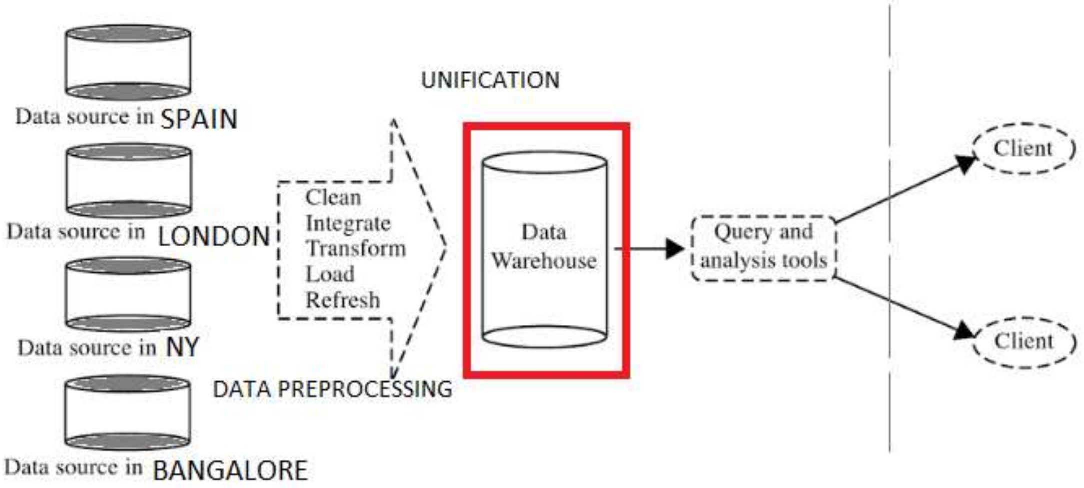

Figure 1 Data Warehouse

A. Process flow in data warehouse

There are four major processes that contribute to a data warehouse:

- Extract and load the data
- Cleaning and transforming the data
- Backup and archive the data
- Managing queries and directing them to the appropriate sources

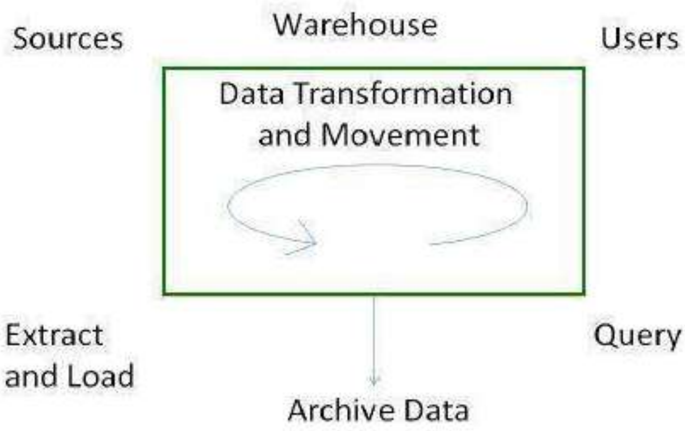

Figure 2 Data transformation and its movement

Figure 2 shows data transformation and its movement. Data extraction takes data from the source systems. Data load takes the extracted data and stores it into the data warehouse. Once the data is extracted and loaded into the temporary

DOI10.17148/IJARCCE.2017.6503

data store, it is time to clean and transform the data. Real world data tends to be incomplete, noisy and inconsistent. Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outlines, and correct inconsistencies in the data. Transforming involves converting the source data into a structure.
In order to recover the data in the event of data loss, software failure or hardware failure, it is necessary to keep regular backups. Archiving involves removing the old data from the system in a format that allows it to be quickly restored whenever required. Query and analysis tools are used in query management process. This process performs the following functions:

- Manages the queries
- Helps speed up the execution time of the queries
- Directs the queries to the most effective system sources.
- Ensures that all system sources are used in the most effective way
- Monitors actual query profile

B. Data Warehousing Architecture
Data warehouses have three-tier architecture:

- Bottom tier
- Middle tier
- Top tier

Figure 3 shows the three tier architecture of a data warehouse.

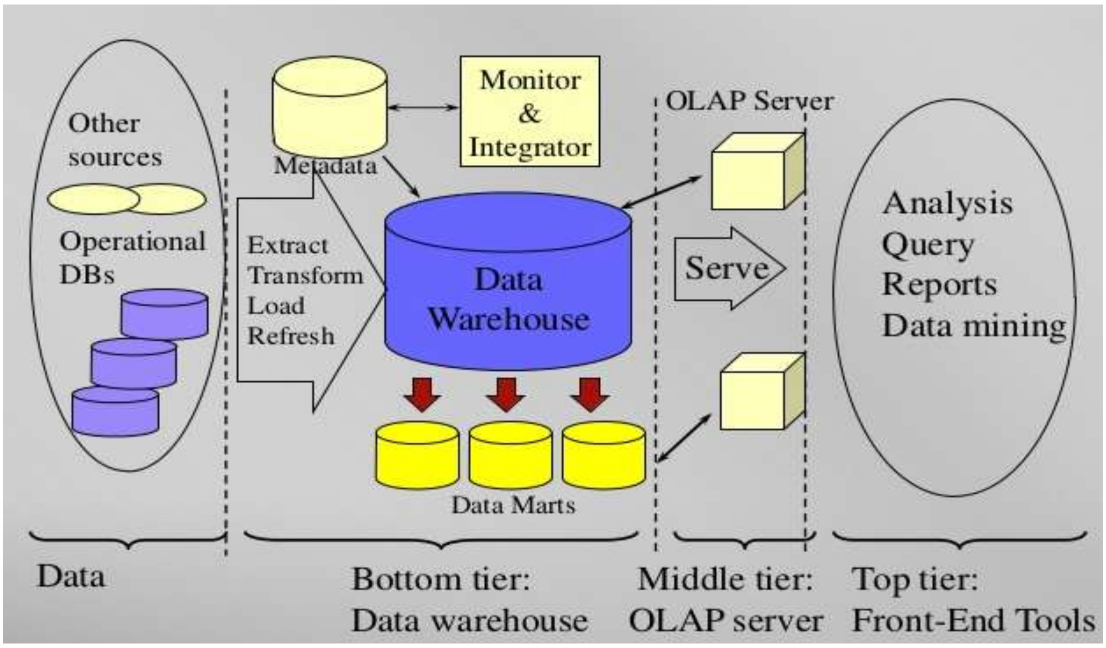

Figure 3 Three tier architecture of a data warehouse

The bottom tier of the architecture consists of the data warehouse server. It is also called the data staging area. It is often the most complex part in the architecture and involves:

- Extraction
- Transformation
- Loading
- Indexing

The middle tier consists of the OLAP server which can be implemented as:

- ROLAP – Relational OLAP maps the operations on multidimensional data to standard relational operations [6]
- MOLAP – Multidimensional OLAP directly implements the multidimensional data and operations. [7]

The middle tier is also called the data presentation area.
The top tier consists of the data access tools. This layer holds:

- Query tools
- Reporting tools

DOI10.17148/IJARCCE.2017.6503

- Analysis tools
- Data Mining tools

C. Data Marts

A data mart is a scaled down version of a data warehouse that focuses on a particular subject area. It is a subset of an organisational data store, usually oriented to a specific purpose or major data subject. Data marts are analytical data stores designed to focus on specific business functions for a specific community within an organization. The reasons for creating a data mart are:

- Easy access to frequently required data
- Improves the end user response time
- To focus on a particular subject area.
- Ease of creation.
- Lower cost than implementing a full data warehouse.

Figure 4 shows data warehouse and its data marts.

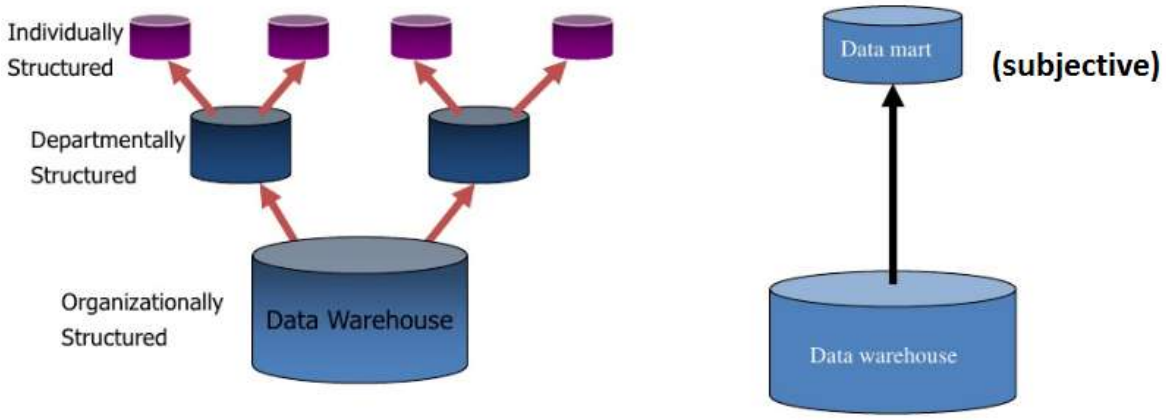

Figure 4 Data warehouse and its data marts

III. APPLICATIONS OF DATA MINING

Data mining, the extraction of hidden predictive information from large databases, is a powerful technology with great potential to help companies focus on the most important information from data warehouses. Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions (KDD). It is a prospective approach rather than the traditional retrospective approach followed by statisticians and other data scientists. Table 1 shows the common real time applications of data mining.

Table 1 Common real time applications of data mining

APPLICATION	DESCRIPTION
Market segmentation	Identifies common characteristics of customers who buy the same products from the same company
Customer churn	Predicts which customers are likely to leave your company and go to a competitor
Fraud detection	Identifies which transactions are most likely to be fraudulent
Direct marketing	Identifies which prospects should be included in a mailing list to obtain highest response rate
Market based analysis	Identifies which products or services are commonly purchased together
Trend analysis	Reveals the difference between a typical customer this month versus last month
Science	Stimulates nuclear explosions; visualises quantum physics
Entertainment	Models customer flow in theme parks; analyses safety of park rides
Insurance and health care	Predicts which customers will buy new policies; identifies behaviour patterns that increase insurance risk; spots fraudulent claims
Manufacturing	Optimises product design, balancing manufacturability and safety; improvises

DOI10.17148/IJARCCE.2017.6503

	shop floor scheduling and machine optimisation
Medicine	Ranks successful therapies for different illnesses; predicts drug efficacy; discovers new drugs and treatments
Oil and gas	Analyses seismic data for signs of underground deposits; prioritises drilling locations; simulates underground flows to improve recovery
Retailing	Discerns buying-behaviour patterns; predicts how customers will react to a marketing campaign

IV. DATA CUBE

Konigtronics Pvt. Ltd would like to keep the sales record with the help of sales data warehouse with respect to time, item, branch and location. Table 2 represents the 2-D view of the sales data of Konigtronics Pvt. Ltd with respect to time, item and a single business location.
Location: Bangalore

Table 2 2-D View of sales data of Konigtronics with respect to time, item and a single location

Time (quarter)	Quarter		
	MQ Sensors	Processors	Monitors
Q1	788	987	765
Q2	678	654	987
Q3	899	875	190
Q4	787	969	908

In this 2-D table, we have records with respect to time and item only. Now let us add one more dimension – location. Say Bangalore, New Delhi and Chennai as shown in table 3.

Table 3 View of sales data of Konigtronics with respect to time, item and three locations

Time	Location: Bangalore			Location: New Delhi			Location: Chennai		
	Item			Item			Item		
	MQ Sensor	Processor	Monitor	MQ Sensor	Processor	Monitor	MQ Sensor	Processor	Monitor
Q1	788	987	765	786	85	987	986	567	875
Q2	678	654	987	659	786	436	980	876	908
Q3	899	875	190	983	909	237	987	100	1089
Q4	787	969	908	537	567	836	837	926	987

The above table can be represented as a 3-D cube as shown in figure 5.

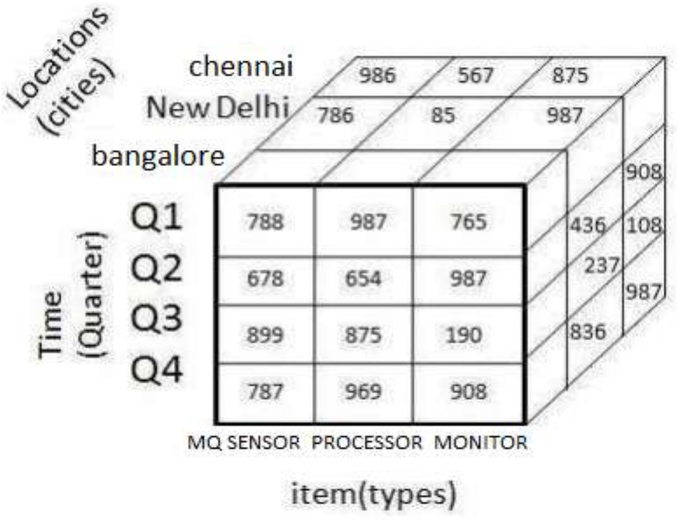

Figure 5 3-D cube

V. CONCLUSION

A data warehouse help businesses to organize, analyze and use their data for decision making. A data warehouse serves as a sole part of a plan-execute-assess “closed-loop” feedback system for the enterprise management. A data warehouse is constructed by integrating data from heterogeneous sources. This integration enhances the effective analysis of data. Also the data in data warehouse is non-volatile, meaning the previous data is not erased when new data is added to it. A data warehouse is kept separate from the operational database and therefore frequent changes in the database are not reflected in the data warehouse.

REFERENCES

[1] Data Mining and Analytics: A Proactive Model - http://www.ijarcce.com/upload/2017/february-17/IJARCCE%20117.pdf
[2] Data Mining, Internet Marketing and Web Mining - http://www.ijarcce.com/upload/2017/march-17/IJARCCE%20117.pdf
[3] Muller H., Freytag J., Problems, Methods, and Challenges in Comprehensive Data Cleansing, Humboldt-Universitatzu Berlin, Germany.
[4] IBM Data Integration - IBM Analytics, www.ibm.com/analytics/us/en/technology/data-integration/
[5] Kimball, R., Caserta, J. The Data Warehouse ETL Toolkit, Wiley and Sons, 2004. ISBN 0-7645-6757-8.
[6] Codd E.F.; Codd S.B. &Salley C.T. (1993)."Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate" (PDF).Codd& Date, Inc. Retrieved 2008-03-05.
[7] "OLAP Council White Paper" (PDF). OLAP Council. 1997. Retrieved 2008-03-18.

BIOGRAPHIES

VISHESH S born on 13th June 1992, hails from Bangalore (Karnataka) and has completed B.E in Telecommunication Engineering from VTU, Belgaum, Karnataka in 2015. He also worked as an intern under Dr Shivananju BN, Department of Instrumentation, IISc, Bangalore. His research interests include Embedded Systems, Wireless Communication and Medical Electronics. He is also the Founder and Managing Director of the company Konigtronics Private Limited. He has guided over a hundred students/lecturers/interns/professionals in their research works and projects. He is also the co-author of many International Research Papers. Presently Konigtronics Private Limited has extended its services in the field of Real Estate, Webpage Designing and Entrepreneurship.

MANU SRINATH hails from Bangalore (Karnataka); he has completed B.E in Telecommunication Engineering from VTU, Belgaum, Karnataka. His research interests include networking, image processing and cryptography. He is the Executive Officer at the company Konigtronics (OPC) Pvt. Ltd.

AKSHATHA C KUMAR from Bangalore (Karnataka) is currently pursuing B.E. in Telecommunication Engineering at B.N.M. Institute of Technology. Her research interests include Data Sciences and Software Engineering.

NANDAN A.S. hails from Bangalore (Karnataka). He is currently pursuing B.E. in Computer Science and Engineering at B.N.M. Institute of Technology. His research interests include Data Sciences and Software Engineering.

DOI10.17148/IJARCCE.2017.6503