Contact
Calea Dorobanților, no. 15-17, room 2017, Bucharest, Romania
Web: http://dbjournal.ro
E-mail: editor@dbjournal.ro

Contents:

Perspectives on Big Data and Big Data Analytics

Elena Geanina ULARU, Florina Camelia PUICAN, Anca APOSTU, Manole VELICANU
[1]Phd. Student, Institute of Doctoral Studies Bucharest
[2]Phd. Student, Institute of Doctoral Studies Bucharest
[3]Phd. Student, Institute of Doctoral Studies Bucharest
[4]Phd. Coordinator, Institute of Doctoral Studies Bucharest
ularugeanina@yahoo.com, puicanflorina@yahoo.com,apostuanca@yahoo.com
manole.velicanu@ie.ase.ro

Nowadays companies are starting to realize the importance of using more data in order to support decision for their strategies. It was said and proved through study cases that "More data usually beats better algorithms". With this statement companies started to realize that they can chose to invest more in processing larger sets of data rather than investing in expensive algorithms. The large quantity of data is better used as a whole because of the possible correlations on a larger amount, correlations that can never be found if the data is analyzed on separate sets or on a smaller set. A larger amount of data gives a better output but also working with it can become a challenge due to processing limitations.
This article intends to define the concept of Big Data and stress the importance of Big Data Analytics.
Keywords: *Big Data, Big Data Analytics, Database, Internet, Hadoop project.*

1 Introduction

Nowadays the Internet represents a big space where great amounts of information are added every day. The IBM Big Data Flood Infographic shows that 2.7 Zettabytes of data exist in the digital universe today. Also according to this study there are 100 Terabytes updated daily through Facebook, and a lot of activity on social networks this leading to an estimate of 35 Zettabytes of data generated annually by 2020. Just to have an idea of the amount of data being generated, one zettabyte (ZB) eguals 10^{21} bytes, meaning 10^{12} GB. [1]

We can associate the importance of Big Data and Big Data Analysis with the society that we live in. Today we are living in an Informational Society and we are moving towards a Knowledge Based Society. In order to extract better knowledge we need a bigger amount of data. The Society of Information is a society where information plays a major role in the economical, cultural and political stage.

In the Knowledge society the competitive advantage is gained through understanding the information and predicting the evolution of facts based on data. The same happens with Big Data. Every organization needs to collect a large set of data in order to support its decision and extract correlations through data analysis as a basis for decisions.

In this article we will define the concept of Big Data, its importance and different perspectives on its use. In addition we will stress the importance of Big Data Analysis and show how the analysis of Big Data will improve decisions in the future.

2. Big Data Concept

The term "Big Data" was first introduced to the computing world by Roger Magoulas from O'Reilly media in 2005 in order to define a great amount of data that traditional data management techniques cannot manage and process due to the complexity and size of this data.

A study on the Evolution of Big Data as a Research and Scientific Topic shows that the term "Big Data" was present in research starting with 1970s but has been

comprised in publications in 2008. [2] Nowadays the Big Data concept is treated from different points of view covering its implications in many fields.
According to MiKE 2.0, the open source standard for Information Management, Big Data is defined by its size, comprising a large, complex and independent collection of data sets, each with the potential to interact. In addition, an important aspect of Big Data is the fact that it cannot be handled with standard data management techniques due to the inconsistency and unpredictability of the possible combinations. [3]
In IBM's view Big Data has four aspects:

Volume: refers to the quantity of data gathered by a company. This data must be used further to obtain important knowledge;
Velocity: refers to the time in which Big Data can be processed. Some activities are very important and need immediate responses, that is why fast processing maximizes efficiency;
Variety: Refers to the type of data that Big Data can comprise. This data can be structured as well as unstructured;
Veracity: refers to the degree in which a leader trusts the used information in order to take decision. So getting the right correlations in Big Data is very important for the business future. [4]
In addition, in Gartner's IT Glosarry Big Data is defined as high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. [5]

According to Ed Dumbill chair at the O'Reilly Strata Conference, Big Data can be described as, "data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it." [6]
In a simpler definition we consider Big Data to be an expression that comprises different data sets of very large, highly complex, unstructured, organized, stored and processed using specific methods and techniques used for business processes.

There are a lot of definitions on Big Data circulating around the world, but we consider that the most important one is the one that each leader gives to its one company's data. The way that Big Data is defined has implication in the strategy of a business. Each leader has to define the concept in order to bring competitive advantage for the company.

The importance of Big Data

The main importance of Big Data consists in the potential to improve efficiency in the context of use a large volume of data, of different type. If Big Data is defined properly and used accordingly, organizations can get a better view on their business therefore leading to efficiency in different areas like sales, improving the manufactured product and so forth.

Big Data can be used effectively in the following areas:

- In information technology in order to improve security and troubleshooting by analyzing the patterns in the existing logs;
- In customer service by using information from call centers in order to get the customer pattern and thus enhance customer satisfaction by customizing services;
- In improving services and products through the use of social media content. By knowing the potential customers preferences the company can modify its product in order to address a larger area of people;
- In the detection of fraud in the online transactions for any industry;
- In risk assessment by analyzing information from the transactions on the financial market.

In the future we propose to anayze the potencial of Big Data and the power that can be enabled through Big Data Analysis.

Big Data challenges

The **understanding of Big Data** is mainly very important. In order to determine the best strategy for a company it is essential that the data that you are counting on must be properly analyzed. Also the time span of this analysis is important because some of them need to be performed very frequent in order to determine fast any change in the business environment.

Another aspect is represented by the **new technologies** that are developed every day. Considering the fact that Big Data is new to the organizations nowadays, it is necessary for these organizations to learn how to use the new developed technologies as soon as they are on the market. This is an important aspect that is going to bring competitive advantage to a business.

The **need for IT specialists** it is also a challenge for Big Data. According to McKinsey's study on Big Data called Big Data: The next frontier for innovation, there is a need for up to 190,000 more workers with analytical expertise and 1.5 million more data-literate managers only in the United States. This statistics are a proof that in order for a company to take the Big Data initiative has to either hire experts or train existing employees on the new field.

Privacy and Security are also important challenges for Big Data. Because Big Data consists in a large amount of complex data, it is very difficult for a company to sort this data on privacy levels and apply the according security. In addition many of the companies nowadays are doing business cross countries and continents and the differences in privacy laws are considerable and have to be taken into consideration when starting the Big Data initiative.

In our opinion for an organization to get competitive advantage from the manipulation of Big Data it has to take very good care of all factors when implementing it. One option of developing a Big Data strategy is presented below. In addition, in order to bring full capabilities to Big Data each company has to take into consideration its own typical business characteristics.

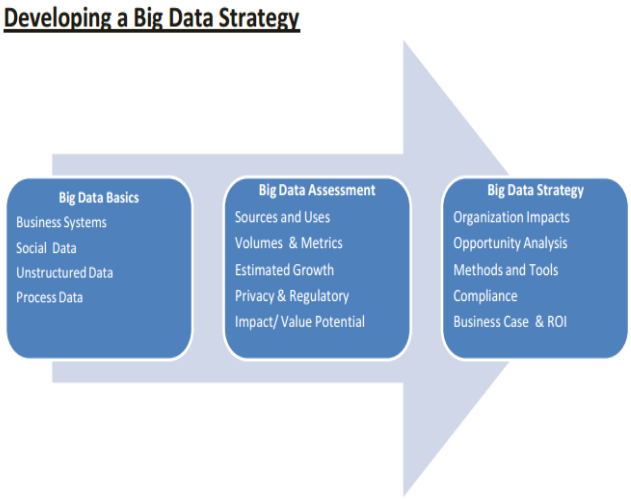


Fig 1. Developing a Big Data Strategy (Source http://www.navint.com) [7]

3 Big Data Analytics

The world today is built on the foundations of data. Lives today are impacted by the ability of the companies to dispose, interrogate and manage data. The development of technology infrastructure is adapted to help generate data, so that all the offered services can be improved as they are used.

As an example, internet today became a huge information-gathering platform due to social media and online services. At any minute they are added data. The explosion of data cannot be any more

measured in gigabytes, since data is bigger there are used etabytes, exabytes, zettabytes and yottabytes.

In order to manage the giant volume of unstructured data stored, it has been emerged the "Big Data" phenomena. It stands to reason that in the commercial sector Big-Data has been adopted more rapidly in data driven industries, such as financial services and telecommunications, which it can be argued, have been experiencing a more rapid growth in data volumes compared to other market sectors, in addition to tighter regulatory requirements and falling profitability. At first, Big Data was seen as a mean to manage to reduce the costs of data management. Now, the companies focus on the value creation potential. In order to benefit from additional insight gained there is the need to assess the analytical and execution capabilities of "Big Data".

To turn big data into a business advantage, businesses have to review the way they manage data within data centre. The data is taken from a multitude of sources, both from within and without the organization. It can include content from videos, social data, documents and machine-generated data, from a variety of applications and platforms. Businesses need a system that is optimised for acquiring, organising and loading this unstructured data into their databases so that it can be effectively rendered and analysed. Data analysis needs to be deep and it needs to be rapid and conducted with business goals in mind.

The scalability of big data solutions within data centres is an essential consideration. Data is vast today, and it is only going to get bigger. If a data centre can only cope with the levels of data expected in the short to medium term, businesses will quickly spend on system refreshes and upgrades. Forward planning and scalability are therefore important.

In order to make every decision as desired there is the need to bring the results of knowledge discovery to the business process and at the same time track any impact in the various dashboards, reports and exception analysis being monitored. New knowledge discovered through analysis may also have a bearing on business strategy, CRM strategy and financial strategy going forward. See figure 2

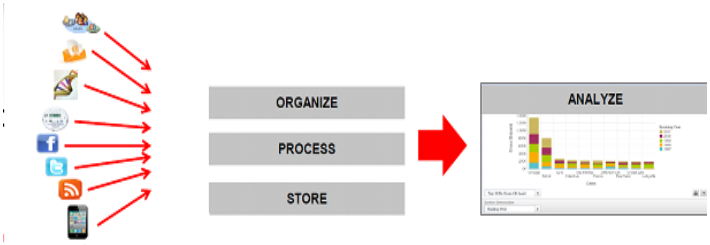


Fig 2. Big Data Management

Up until mid 2009 ago, the data management landscape was simple: Online transaction processing (OLTP) systems (especially databases) supported the enterprise's business processes; operational data stores (ODSs) accumulated the business transactions to support operational reporting, and enterprise data warehouses (EDWs) accumulated and transformed business transactions to support both operational and strategic decision making.

Big Data Management is based on capturing and organizing relevant data. Data analytics supposes to understand that happened, why and predict what will happen. A deeper analytics means new analytical methods for deeper insights.[9]

Big data analytics and the Apache Hadoop open source project are rapidly emerging as the preferred solution to business and technology trends that are disrupting the traditional data management and processing landscape. Enterprises can gain a competitive advantage by being early adopters of big data analytics. Even though big data analytics can be technically challenging, enterprises should not delay

implementation. As the Hadoop projects mature and business intelligence (BI) tool support improves, big data analytics implementation complexity will reduce, but the early adopter competitive advantage will also wane. Technology implementation risk can be reduced by adapting existing architectural principles and patterns to the new technology and changing requirements rather than rejecting them. [10]

Big data analytics can be differentiated from traditional data-processing architectures along a number of dimensions:

- Speed of decision making being very important for decision makers
- Processing complexity because it eases the decision making process
- Transactional data volumes which are very large
- Data structure data can be structured and unstructured
- Flexibility of processing/analysis consisting in the amount of analysis that can be performed on it
- Concurrency [9]

The big data analytics initiative should be a joint project involving both IT and business. IT should be responsible for deploying the right big data analysis tools and implementing sound data management practices. Both groups should understand that success will be measured by the value added by business improvements that are brought about by the initiative.

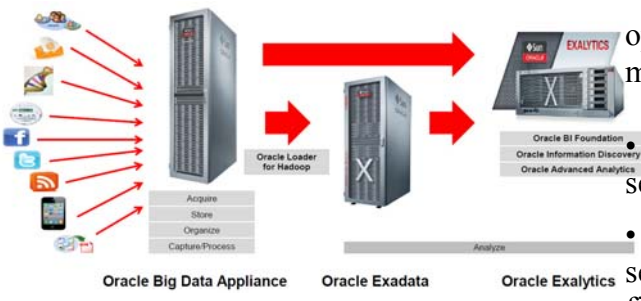


Fig 3. Oracle Big Data Solution (Source: myoracle.com)

In terms of Big Data Management and analytics Oracle is offering Engineered Systems as Big Data Solutions (Fig.3), such as Oracle Big Data Appliance, Oracle Exadata and Oracle Exalytics. Big Data solutions combine best tools for each part of the problem. The traditional business intelligence tools rely on relational databases for storage and query execution and did not target Hadoop. Oracle BI combined with Oracle Big Data Connectors. The architecture supposes to load key elements of information from Big Data sources into DBMS. Oracle Big Data connectors, Hive and Hadoop aware ETL such as ODI provide the needed data integration capabilities. The key benefits are that the business intelligence investments and skills that are leveraged, there are made insights from Big Data consumable for business users, there are combined Big Data with Application and OLTP data. [11]

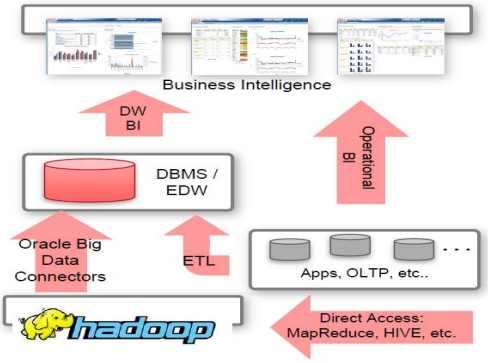


Fig 4. BI and Data Warehousing on Big Data (Source: myoracle.com)

Big Data provides many opportunities for deep insights via data mining:

• Uncover relationships between social sentiment and sales data

• Predict product issues based on diagnostic sensor data generated by products in the field

• In fact, the signal-to-noise issues often mean deep analytics to mine insight hidden in the noise is essential, as many forms of Big Data are simply not consumable in raw form

"Big Data" is a Data Management & Analytics market opportunity driven by new market requirements. In-Database Analytics – Data Mining there are used Big Data Connectors to combine Hadoop and DBMS data for deep analytics. Also there is the need to re-use SQL skills to apply deeper data mining techniques or re-use skills for statistical analysis. Everything is all about "Big Data" instead of RAM-scale data. This is how the predictive learning of relationships between knowledge concepts and business events is done. [12]

Big-Data presents a significant opportunity to create new value from giant data. It is important to determine appropriate governance procedures in order to manage development and implementations over the life of the technology and data. Failure to consider the longer term implications of development will lead to productivity issues and cost escalations.

On the face of it, the cost of physically storing large quantities of data is dramatically reduced by the simplicity by which data can be loaded into a Big-Data cluster because there is no longer required a complex ETL layer seen in any more traditional Data Warehouse solutions. The cluster itself is also typically built using low cost commodity hardware and analysts are free to write code in almost any contemporary language through the streaming API available in Hadoop.

- The business logic used within an ETL flow to tokenise a stream of data and apply data quality standards to it must be encoded (typically using Java) within each Map-Reduce program that processes the data and any changes in source syntax or semantics [8]
- Although the storage nodes in a Hadoop cluster may be built using low cost commodity x86 servers, the master nodes (Name Node, Secondary Name Node and Job Tracker) requiring higher resilience levels to be built into the servers if disaster is to be avoided. Map-Reduce operations also generate a lot of network chatter so a fast private network is recommended. These requirements combine to add significant cost to a production cluster used in a commercial setting. [8]
- Compression capabilities in Hadoop are limited because of the HDFS block structure and require an understanding of the data and compression technology to implement adding to implementation complexity with limited impact on storage volumes.

  Other aspects to consider include the true cost of ownership of pre-production and production clusters such as the design build and maintenance of the clusters themselves, the transition to production of Map-Reduce code to the production cluster in accordance with standard operational procedures and the development of these procedures. [8]

Whatever the true cost of Big-Data compared to a relational data storage approach, it is important that the development of Big-Data strategy is consciously done, understanding the true nature of the costs and complexity of the infrastructure, practice and procedures that are put in place.

4 Big Data Analytics Software

Currently, the trend is for enterprises to re-evaluate their approach on data storage, management and analytics, as the volume and complexity of data is growing so rapidly and unstructured data accounting is for 90% of the data today.

Every day, 2.5 quintillion bytes of data are created — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from various sources such as: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals, web and software logs, cameras, information-sensing mobile devices, aerial sensory technologies and genomics. This data is referred to as ***big data.***

"Legacy systems will remain necessary for specific high-value, low-volume workloads, and compliment the use of Hadoop - optimizing the data management structure in the organization by putting the right Big Data workloads in the right systems"[14].

As it was mentioned in the Introduction Big data spans four dimensions: Volume, Velocity, Variety, and Veracity

- *Volume:* Enterprises are awash with ever-growing data of all types, easily amassing terabytes - even petabytes - of information(e.g. turn 12 terabytes of Tweets created each day into improved product sentiment analysis; convert 350 billion annual meter readings to better predict power consumption);
- *Velocity:* For time-sensitive processes such as catching fraud, big data flows must be analysed and used as they stream into the organizations in order to maximize the value of the information(e.g. scrutinize 5 million trade events created each day to identify potential fraud; analyze 500 million daily call detail records in real-time to predict customer churn faster).
- *Variety:* Big data consists in any type of data - structured and unstructured data such as text, sensor data, audio, video, click streams, log files and more. The analysis of combined data types brings new aspect for problems, situations etc.(e.g. monitor 100's of live video feeds from surveillance cameras to target points of interest; exploit the 80% data growth in images, video and documents to improve customer satisfaction);
- *Veracity:* Since one of three business leaders don't trust the information they use to make decisions, establishing trust in big data presents a huge challenge as the variety and number of sources grows.

Apache **Hadoop** is a fast-growing big-data processing platform defined as "an open source software project that enables the distributed processing of large data sets across clusters of commodity servers"[15]. It is designed to scale up from a single server to thousands of machines, with a very high degree of fault tolerance.

Rather than relying on high-end hardware, the resiliency of these clusters comes from the software's ability to detect and handle failures at the application layer.

Developed by Doug Cutting, Cloudera's Chief Architect and the Chairman of the Apache Software Foundation, Apache Hadoop was born out of necessity as data from the web exploded, and grew far beyond the ability of traditional systems to handle it. Hadoop was initially inspired by papers published by Google outlining its approach to handling an avalanche of data, and has since become the de facto standard for storing, processing and analyzing hundreds of terabytes, and even petabytes of data.

Apache Hadoop is 100% open source, and pioneered a fundamentally new way of storing and processing data. Instead of relying on expensive, proprietary hardware and different systems to store and process data, Hadoop enables distributed parallel processing of huge amounts of data across inexpensive, industry-standard

servers that both store and process the data, and can scale without limits.

In today's hyper-connected world where more and more data is being created every day, Hadoop's breakthrough advantages mean that businesses and organizations can now find value in data that was recently considered useless.

Hadoop can handle all types of data from disparate systems: structured, unstructured, log files, pictures, audio files, communications records, email - regardless of its native format. Even when different types of data have been stored in unrelated systems, it is possible to store it all into Hadoop cluster with no prior need for a schema.

By making all data useable, Hadoop provides the support to determine inedited relationships and reveal answers that have always been just out of reach.

In addition, Hadoop's cost advantages over legacy systems redefine the economics of data. Legacy systems, while fine for certain workloads, simply were not engineered with the needs of Big Data in mind and are far too expensive to be used for general purpose with today's largest data sets.

Apache Hadoop has two main subprojects:

- *MapReduce* - The framework that understands and assigns work to the nodes in a cluster. Has been defined by Google in 2004 and is able to distribute data workloads across thousands of nodes. It is based on "break problem up into smaller sub-problems" strategy and can be exposed via SQL and in SQL-based BI tools;
- *Hadoop Distributed File System (HDFS)* - An Apache open source distributed file system that spans all the nodes in a Hadoop cluster for data storage. It links together the file systems on many local nodes to make them into one big file system. HDFS assumes nodes will fail, so it achieves reliability by replicating data across multiple nodes

HDFS is expected to run on high-performance commodity hardware; it is known for highly scalable storage and automatic data replication across three nodes for fault tolerance. Furthermore, automatic data replication across three nodes eliminates need for backup (write once, read many times).

Hadoop is supplemented by an ecosystem of Apache projects, such as Pig, Hive and Zookeeper, that extend the value of Hadoop and improve its usability. Due to the cost-effectiveness, scalability and streamlined architectures, Hadoop changes the economics and the dynamics of large scale computing, having a remarkable influence based on four salient characteristics. Hadoop enables a computing solution that is:

- *Scalable:* New nodes can be added as needed, and added without needing to change data formats, how data is loaded, how jobs are written, or the applications on top.
- *Cost effective:* Hadoop brings massively parallel computing to commodity servers. The result is a sizeable decrease in the cost per terabyte of storage, which in turn makes it affordable to model all your data.
- *Flexible:* Hadoop is schema-less, and can absorb any type of data, structured or not, from any number of sources. Data from multiple sources can be joined and aggregated in arbitrary ways enabling deeper analyses than any one system can provide.
- *Fault tolerant:* When you lose a node, the system redirects work to another location of the data and continues processing without missing a beat.

Text mining makes sense of text-rich information such as insurance claims, warranty claims, customer surveys, or the growing streams of customer comments on social networks.

Optimization helps retailers and consumer goods makers, among others, with tasks such as setting prices for the best possible balance of strong-yet-profitable sales. Forecasting is used by insurance companies, for example, to estimate exposure or losses in the event of a hurricane or flood.

Cost will certainly be a software selection factor as that's a big reason companies are adopting Hadoop; they're trying to retain and make use of all their data, and they're expecting cost savings over conventional relational databases when scaling out over hundreds of Terabytes or more. Sears, for example, has more than 2 petabytes of data on hand, and until it implemented Hadoop two years ago, Shelley says the company was constantly outgrowing databases and still couldn't store everything on one platform.

Once the application can run on Hadoop it will presumably be able to handle projects with even bigger and more varied data sets, and users will be able to quickly analyze new data sets without the delays associated with transforming data to meet a rigid, predefined data model as required in relational environments.

From architectural point of view, Hadoop consists of the Hadoop Common which provides access to the filesystems supported by Hadoop. The Hadoop Common package contains the necessary JAR files and scripts needed to start Hadoop. The package also provides source code, documentation, and a contribution section which includes projects from the Hadoop Community.

For effective scheduling of work, every Hadoop-compatible filesystem should provide location awareness: the name of the rack (more precisely, of the network switch) where a worker node is. Hadoop applications can use this information to run work on the node where the data is, and, failing that, on the same rack/switch, reducing backbone traffic. The Hadoop Distributed File System (HDFS) uses this when replicating data, to try to keep different copies of the data on different racks. The goal is to reduce the impact of a rack power outage or switch failure so that even if these events occur, the data may still be readable.

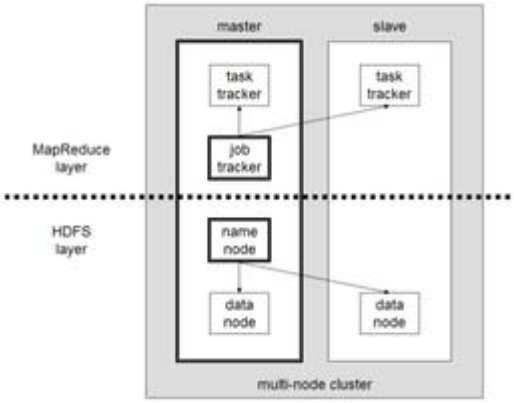


Fig 5. A multi-node Hadoop cluster[13]

As shown in Fig. 5, a small Hadoop cluster will include a single master and multiple worker nodes. The master node consists of a JobTracker, TaskTracker, NameNode, and DataNode.

A slave or *worker node* acts as both a DataNode and TaskTracker, though it is possible to have data-only worker nodes, and compute-only worker nodes; these are normally only used in non-standard applications.

Hadoop requires JRE 1.6 or higher. The standard startup and shutdown scripts require Secure Shell(SSH) to be set up between nodes in the cluster.

In a larger cluster, the HDFS is managed through a dedicated NameNode server to host the filesystem index, and a secondary NameNode that can generate snapshots of the namenode's memory structures, thus preventing filesystem corruption and reducing loss of data.

Similarly, a standalone JobTracker server can manage job scheduling.

In clusters where the Hadoop MapReduce engine is deployed against an alternate filesystem, the NameNode, secondary NameNode and DataNode architecture of HDFS is replaced by the filesystem-specific equivalent.

One of the cost advantages of Hadoop is that because it relies in an internally redundant data structure and is deployed on industry standard servers rather than expensive specialized data storage systems, you can afford to store data not previously viable.

Big data is more than simply a matter of size; it is an opportunity to find insights in new and emerging types of data and content, to make businesses more agile and to answer questions that were previously considered beyond reach.

Enterprises who build their Big Data solution can afford to store literally all the data in their organization, and keep it all online for real-time interactive querying, business intelligence, analysis and visualization.

5 Conclusions

The year 2012 is the year when companies are starting to orient themselves towards the use of Big Data. That is why this articol presents the Big Data concept and the technologies associated in order to understand better the multiple beneficies of this new concept ant technology.

In the future we propose for our research to further investigate the practical advantages that can be gain through Hadoop.

References

[1] G. Noseworthy, Infographic: Managing the Big Flood of Big Data in Digital Marketing, 2012 http://analyzingmedia.com/2012/infographic-big-flood-of-big-data-in-digital-marketing/

[2] H. Moed, The Evolution of Big Data as a Research and Scientific Topic: Overview of the Literature, 2012, ResearchTrends, *http://www.researchtrends.com*

[3] MIKE 2.0, Big Data Definition, http://mike2.openmethodology.org/wiki/Big_Data_Definition

[4] P. Zikipoulos, T. Deutsch, D. Deroos, Harness the Power of Big Data, 2012, http://www.ibmbigdatahub.com/blog/harness-power-big-data-book-excerpt

[5] Gartner, Big Data Definition, http://www.gartner.com/it-glossary/big-data/

[6]E. Dumhill, "What is big data?", 2012 ,http://strata.oreilly.com/2012/01/what-is-big-data.html

[7] A Navint Partners White Paper, "Why is BIG Data Important?" May 2012, http://www.navint.com/images/Big.Data.pdf

[8] Greenplum. A unified engine for RDBMS and Map Reduce, 2009. http://www.greenplum.com/resources/mapreduce/.

[9] For Big Data Analytics There's No Such Thing as Too Big The Compelling Economics and Technology of Big Data Computing, White Paper, March 2012, By: 4syth.com, Emerging big data thought leaders

[10] Big data: The next frontier for innovation, competition, and productivity. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. McKinsey Global Institute. May 2011

[11] Oracle Information Architecture: An Architect's Guide to Big Data, An Oracle White Paper in Enterprise Architecture August 2012

[12] http://bigdataarchitecture.com/

[13]http://www.oracle.com/us/corporate/press/1453796

[14]http://www.informationweek.com/softw

are/business-intelligence/sas-gets-hip-to-hadoop-for-big-data/240009035?pgno=2

[15]http://en.wikipedia.org/wiki/Apache_Hadoop

Elena-Geanina ULARU graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Academy of Economic Studies in 2008. She holds a Master Degree obtained at Faculty of Cybernetics, Statistics and Economic Informatics of the Academy of Economic Studies at the Academy of Economic Studies and is currently a Phd. Student, in the second year, at the Institute of Doctoral Studies, doing her reasarch at the University of Economics from Prague.

Florina Camelia PUICAN is a Phd. Student, in the second year, at the Institute of Doctoral Studies. Bucharest. In 2008, she graduated from Faculty of Business Administration with teaching in foreign languages (English), at the Academy of Economic Studies, Bucharest and in 2009, from Faculty of Mathematics and Computer Science, section Computer Science, University of Bucharest. From 2010, she holds a Master Degree obtained at Faculty of Business Administration with teaching in foreign language (English), at the Academy of Economic Studies, Bucharest. During her studies and work experience she undertook a wide range of skills in economics, information technology and information systems for business, design and management of information systems and databases.

Anca Apostu has graduated The Academy of Economic Studies from Bucharest (Romania), Faculty of Cybernetics, Statistics and Economic Informatics in 2006. She has a Master diploma in Economic Informatics from 2010 and in present she is a Ph.D. Candidate in Economic Informatics with the Doctor's Degree Thesis: "Informatics solution in a distributed environment regarding unitary tracking of prices". Her scientific fields of interest include: Economics, Databases, Programming, Information Systems, Information Security and Distributed Systems.

Manole VELICANU is a Professor at the Economic Informatics, Cybernetics and Statistics Department at the Faculty of Cybernetics, Statistics and Economic Informatics from the Academy of Economic Studies of Bucharest. He has graduated the Economic Informatics from the Academy of Economic Studies in 1976, holds a PhD diploma in Economic Informatics from 1994 and starting with 2002 he is a PhD coordinator in the field of Economic Informatics. He is the author of 21 books in the domain of economic informatics, 84 published articles, 74 scientific papers presented at conferences. He participated (as director or as team member) in more than 50 research projects that have been financed from national and international research programs. His fields of interest include: Database Systems, Business Intelligence, Information Systems, Artificial Intelligence, Programming languages.

Real-Time Business Intelligence for the Utilities Industry

Janina POPEANGĂ, Ion LUNGU
Academy of Economic Studies, Bucharest, Romania,
janina.popeanga@yahoo.com; ion.lungu@ie.ase.ro

In today's competitive environment with rapid innovation in smart metering and smart grids, there is an increased need for real-time business intelligence (RTBI) in the utilities industry. Giving the fact that this industry is an environment where decisions are time sensitive, RTBI solutions will help utilities improve customer experiences and operational efficiencies.
The focus of this paper is on the importance of real-time business intelligence (RTBI) in the utilities industry, outlining our vision of real-time business intelligence for this industry. Besides the analysis in this area, the article presents as a case study the Oracle Business Intelligence solution for utilities.
Keywords: *real-time business intelligence, data latency, external real-time data cache, real-time data warehouse, Oracle Utilities Business Intelligence*

1 Introduction

Over the past years, energy consumption has increased dramatically. The rise in consumption of energy resources has meant increasing costs and CO_2 emissions, and the reduction of non-renewable supplies.
In an effort to cut costs and CO_2 emissions, measures are being taken to reduce energy consumption and to use more energy from renewable sources.
The term smart grid has been frequently used in the last few years in order to meet the challenges facing developed and developing countries alike, such as the growing demand for electric power, the need to increase efficiency in energy conversion, delivery, consumption, the provision of high quality power, and the integration of renewable resources for sustainable development. [1]
The smart grid will provide a large volume of sensor and meter data that will require intelligent analytics that move further than data management, querying and reporting.
In addition to increased amounts of data, utilities must manage an increasingly varied set of data from a variety of sources. Sources can include real-time data from external resources such as weather and geospatial information or real-time data about energy production and consumption. Integrating these sources of data in order to make real-time business decisions will require more advanced business solutions.

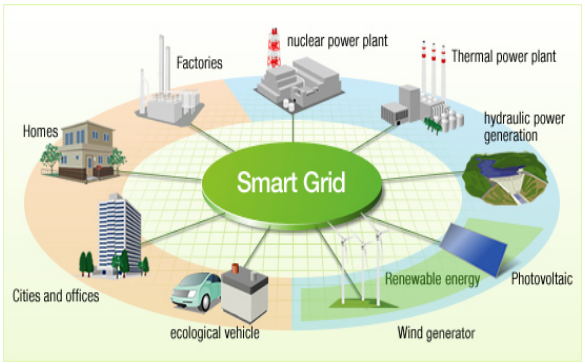


Fig. 1. Smart grid [2]

To survive in this competitive world, a utility enterprise must have the ability to integrate data from numerous sources, compile and filter that data and also analyze and present the data in a clear way in order to support rapid and confident decisions.
True business intelligence can support, grow and ensure the success of the utilities enterprise, and assist the business user in gathering and analyzing data that is critical to forecast and predict demand, anticipate pricing adjustments, manage utility programs, and customers, installations, equipment, facilities, billing, and industry and governmental regulations. [3]

But utilities industry is an environment where decisions are time sensitive, so this paper focuses on the importance of real-time business intelligence for utilities.
The rest of the paper is structured as follows. Sections 2 and 3 define business intelligence and real-time business intelligence. Section 4 makes a review of relevant literature. Section 5 presents Oracle Utilities Business Intelligence solution, while section 6 outlines our vision of real-time business intelligence for the utilities industry. Finally, we conclude this paper in section 7 with a number of results and observations.

2. Defining Business Intelligence

There are numerous definitions of business intelligence by some professionals in the industry.
According to Wayne W. Eckerson, Director of Research and Services for The Data Warehousing Institute (TDWI), "business intelligence is an umbrella term that encompasses a raft of data warehousing, and data integration technologies as well as querying, reporting and analysis tools that fulfill the promise of giving business users self-service access to information". [4]
Cindi Howson thinks that "business intelligence allows people at all levels of an organization to access, interact with, and analyze data to manage the business, improve performance, discover opportunities and operate efficiently."
Common functions of business intelligence technologies are reporting, online analytical processing, analytics, data mining, process mining, complex event processing, business performance management, benchmarking, text mining, predictive analytics and prescriptive analytics.

Fig. 2. Some of the general functions of BI systems [5]

Powerful BI features include personalised dashboards, automated alerts, graphs, charts, gauges and other view options that enable clear, concise display of data with complete drill through analytical capability. [3]
Business intelligence solutions help energy companies in many ways, by enabling them to [6]:

- Optimize the supply chain by providing data access to suppliers, distributors, and customers to enhance performance and responsiveness (all while reducing costs);
- Improve stock control by providing visibility across the organization and supply chain to enhance just-in-time management and reduce excess inventory;
- Minimize procurement inefficiencies by analyzing supplier performance, and driving negotiations and pricing structures;
- Respond quickly to market opportunities by tracking and analyzing operational data from inventory, financial, point-of-sale, and marketing;
- Differentiate and refine product offering by analyzing historical information and assessing product profitability on a geographic basis;

- Strengthen customer relationships and increase their value by tracking customer behavior and service issues, better targeting promotions, and improving service deliver.

According to Dr. Richard Hackathorn, creator of the Time-Value Curve, "the value of data is directly proportionate to how fast a business can react to it. In other words, a corporation loses money every time it delays getting information into the hands of decision-makers." [7]
Latency is the temporal delay between the moment of an event initiation and the moment an action is taken to respond to that event.

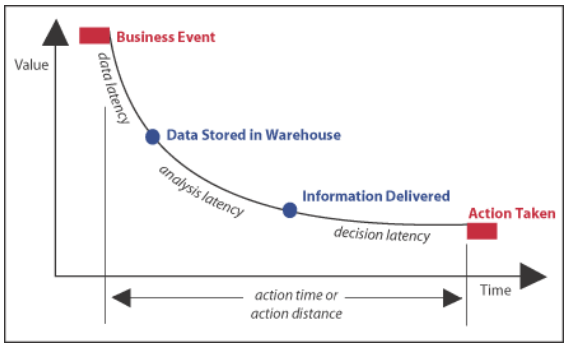


Fig. 3. Latency into the decision-making process [8]

There are three types of latency in a decision making process [9]:
• *data latency* - the period of time needed to collect the data from the source systems, to prepare it for analysis and save it into the data warehouse or data centers;
• *analytic latency* - the period of time needed to access and analyze the data, to transform the data in information, to apply the business rules.
• *decisional latency* - the period of time needed to review the analysis, decide the action to be taken and implement the action.
The degree of latency in a BI system is one of the most important issues because business executives and analytics simply want these systems to deliver the right information in the right format and to the right people, at the right time, so they can make optimal business decisions. (**Fig. 4**)

Fig. 4. The role of BI systems

3. Defining Real-Time Business Intelligence

Real-time business intelligence is the process of delivering information about business operations as they occur, with minimum latency.
All real-time BI systems have some latency, but the goal is to minimize the time from the business event happening to a corrective action or notification being initiated. [10]
Right-time is a better term to use than real time. "Right-time implies that different business situations and events require different response or action times. When planning a right-time processing environment, it is important to match technology requirements to the actual action times required by the business - some situations require close to a real-time action, whereas with others, a delay of a few minutes or hours is acceptable." [11]

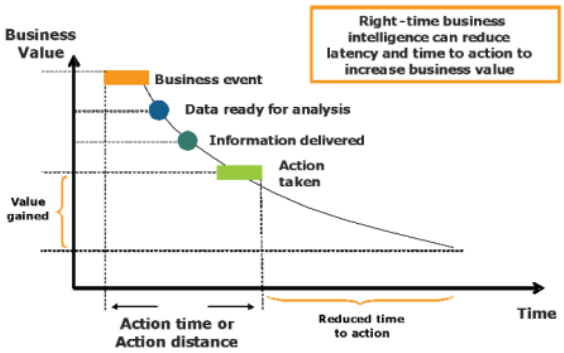


Fig. 5. RTBI latency [12]

In this context, real-time means a range from milliseconds to a few seconds after the business event has occurred.

For example, in order to keep the smart grid performing optimally, a system for monitoring and managing an electrical utility needs data delivered in milliseconds, so that problems can be solved within seconds or minutes. In this example, data must be accessed, processed and delivered in milliseconds. Besides data latency, data unavailability is another impediment to efficient real-time business intelligence.

Giving the fact that utilities are dependent on real-time business intelligence, the unavailability of this intelligence due to a failed system could stop the operations. Great availability of the real-time business intelligence services is vital.

RTBI provides almost the same functionality as the traditional business intelligence, but operates on data that is extracted from operational data sources with zero latency, and provides means to propagate actions back into business processes in real-time. [13]

While traditional business intelligence presents historical data for manual analysis, real-time business intelligence compares current business events with historical patterns to detect problems or opportunities automatically. This automated analysis capability enables corrective actions to be initiated and or business rules to be adjusted to optimize business processes. [14]

A real-time business intelligence system is based on a real-time ETL and a real-time data warehouse.

Also known as active data warehouse, real-time data warehouse is a combination of fast technologies and fast-paced business processes.

One of the most challenging parts of building any data warehouse is the process of extracting, transforming and loading (ETL) the data from the sources. Talking about real-time data warehouse, additional challenges are being introduced. The traditional ETL process involves downtime of the data warehouse while the loads are performed, but when loading real-time data, there cannot be any downtime of the system. For some applications, increasing the frequency of the current data load may be a solution.

Beside the raised costs, real-time data warehouse brings up some important issues like data modeling, scalability, OLAP and query tools. To resolve these problems, researchers proposed various solutions.

Storing real-time data along with the historical data, in the same fact tables, can be a good idea from the data modeling perspective because a real-time data warehouse is modeled just like a traditional data warehouse. But many query and OLAP tools that are not so real-time aware and tables frequently changings bring up another issue – caching.

From the administration perspective, storing real-time data in separate fact tables is the most complex approach.

Storing real-time data in different tables from historical data, but in the same table structure, and using database views to combine these smaller tables, more easily updated, into a single logical table do not resolve caching issues. Nobody wants that the query tools return old cache results when we need real-time data.

4. Literature review

The utilities industry is expected to be one of the fastest-growing industries in adopting business intelligence technology in the next few years.

IDC Energy Insights announced in 2011 the availability of the new report "Business Strategy: Ready for the Dip…Err…Plunge? Utility Business Analytics", that exposes the critical need for the utilities industry to use business intelligence solutions to support automation processes, to take better decisions and to allow customers to manage their energy lifestyles.

Unlike any other research available in the industry, this report details how utilities can efficiently leverage business analytics in both the near and long term to improve operations, increase customer satisfaction, and continuously optimize business decisions. [15]

In its "2012 Utility Industry Survey on Business Intelligence/Analytics," BRIDGE Energy Group questioned more than 14.000 energy industry officers on their experiences with business intelligence solutions.

29 percent of utilities reported that they are planning a major business intelligence program within the next two years, and another 62 percent were planning smaller scale projects, like adding a predictive analytics tool. [16]

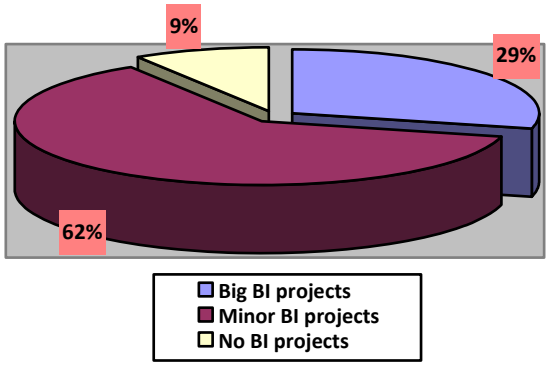


Fig. 6. Utilities expectations

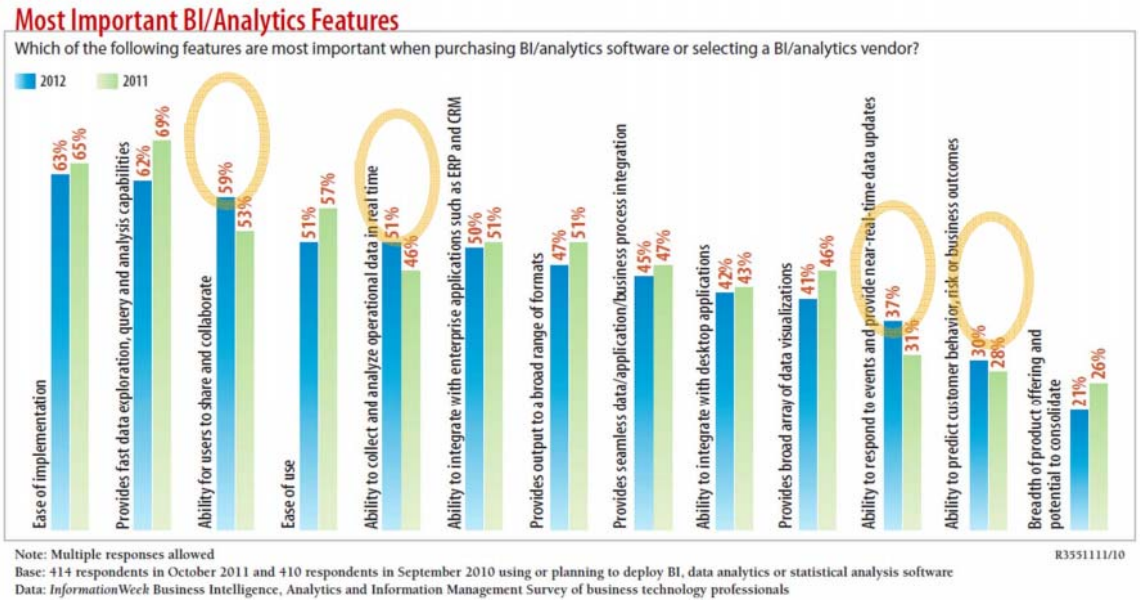


Fig. 7. Most important BI features [17]

According to numerous surveys, utilities expect to expand business intelligence with more programs that involve predictive modeling, the ability to collect and analyze operational data in real-time, and to respond to events and provide near-real-time data updates. (**Fig.7**)

In the academic literature, few articles deal with real-time business intelligence:

- ✓ In "Towards real-time business intelligence" article, B. Azvine et. al. [18] discuss the issues and problems of current BI systems and then outlines their vision of future real-time BI.
- ✓ In 2006, B. Azvine et al. [13] present their vision of what future real-time business intelligence would provide, discuss the technology challenges involved, and describe some of their programmes towards the implementation of our vision.
- ✓ B.S. Sahay and Jayanthi Ranjan [10] focus on the necessity of real-time BI in supply chain analytics. They believe that supply chain analytics using real time BI in organizations will derive better operational efficiency and KPI for any organization in SCM.
- ✓ D. Sandu [9] reviews the differences in the various types of business intelligence.
- ✓ Judith. R. Davis [19] presents a case study which describes the specific business problem, the right-time BI solution, benefits achieved, return on investment (ROI), features critical to success and implementation advice.

5. Oracle Utilities Business Intelligence – case study

Oracle Utilities helps utilities prepare for smart metering and smart grid initiatives that enhance efficiency and provide critical intelligence metrics that can help drive more-informed energy and water usage decisions for consumers and businesses. [20]

Oracle Business Intelligence for Utilities facilitates utilities to simply organize data into reports, ad-hoc queries, in-depth analyses and set up notifications and alerts. This solution delivers intelligence thru maps, charts and graphics, making it easy to manage complex data and to understand relationships. Utilities can use these ***near real-time*** visuals to improve decision-making and to update rapidly on situations like outage restoration.

The Oracle Utilities Business Intelligence data-warehouse is a separate database from your operational database. All data extracted from the production system and transferred to the Oracle Utilities Business Intelligence data-warehouse is held in star schemas.[20] (**Fig. 8**)

The tables in a star-schema are divided into two categories: facts and dimensions. Every star-schema has a single fact table (at the center of the star) and one or more dimensions [20]:

• *Fact tables* contain individual rows for every occurrence of a fact in the production system. Fact tables contain columns called measures. It is these columns that are aggregated to calculate key performance indicators (KPIs).

• *Dimension tables* are used to "slice" the facts in different ways. For example, the star schema above would allow users to "slice" the financial fact by the attributes on the 6 dimensions linked to it.

ETL programs are provided for every fact and dimension in Oracle Utilities Business Intelligence. (**Fig. 9**)

The extract programs are extracting operational data and performing some transformation activities. A separate extract program is used for every distinct fact and dimension.

Figure 10 illustrates the components involved in Oracle Utilities Business Intelligence's ETL methodology.

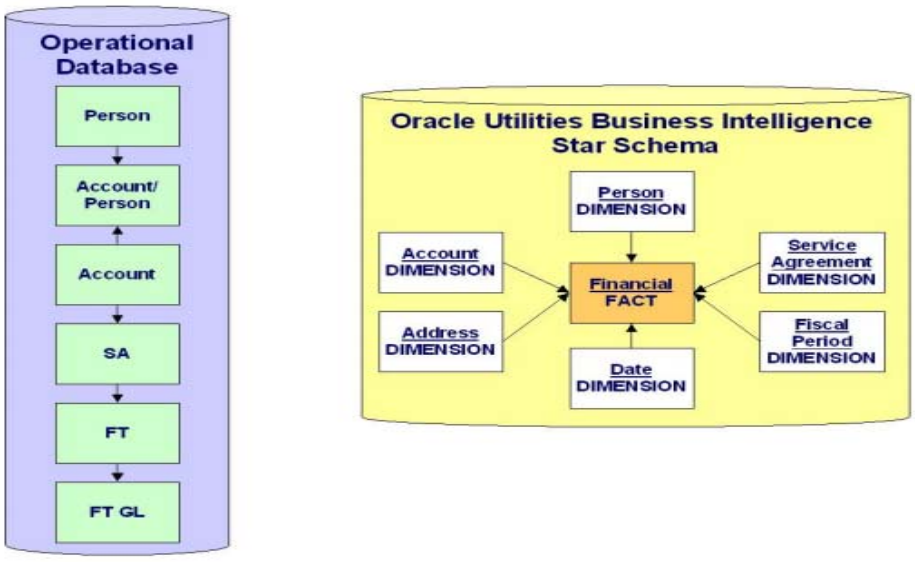


Fig.8. Oracle Utilities Business Intelligence – star schema [20]

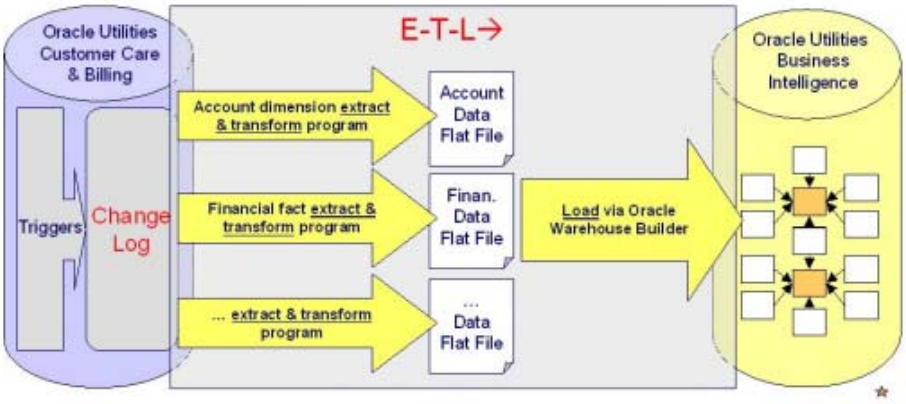


Fig.9. ETL process [20]

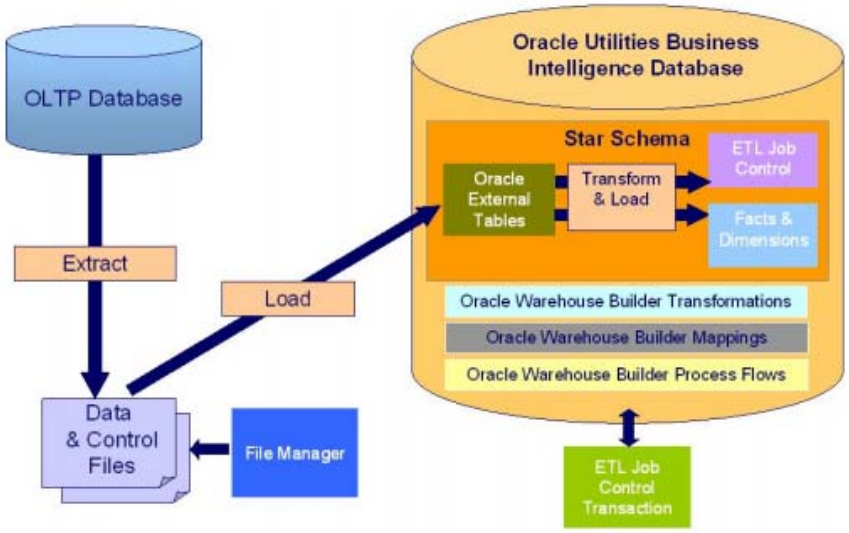


Fig.10. ETL methodology [20]

6. Our RTBI vision for the utilities industry

This paper proposes a RTBI solution for the utilities industry, considering the type of data source.

For real-time data, such as weather data, data about sensors and meters states, production and consumption data, we propose to use an external real-time data cache. This data cache contains only the tables that are real-time, while non-real-time date are extracted, transformed and loaded directly into the traditional data warehouse.

Using an external real-time data cache we can eliminate performance problems associated with the process of integrating real-time data into a data warehouse.

Also, this can solve other problems like internal inconsistency and data latency.

Having all the real-time activity on the external cache database, the warehouse does not support any additional load, so the scalability and query problems will be solved.

The connection between the real-time data cache and data warehouse is accomplished by scheduled and real-time updates when certain conditions are met. For example, production and consumption data are updated hourly, while alerts based on measurable factors or alerts caused by the malfunctioning equipment need to be updated instantly into the data warehouse.

In order to create advanced business analytics, with a just-in-time information merging solution, we can easily merge real-time data from the external real-time data cache with the historical information from the data warehouse.

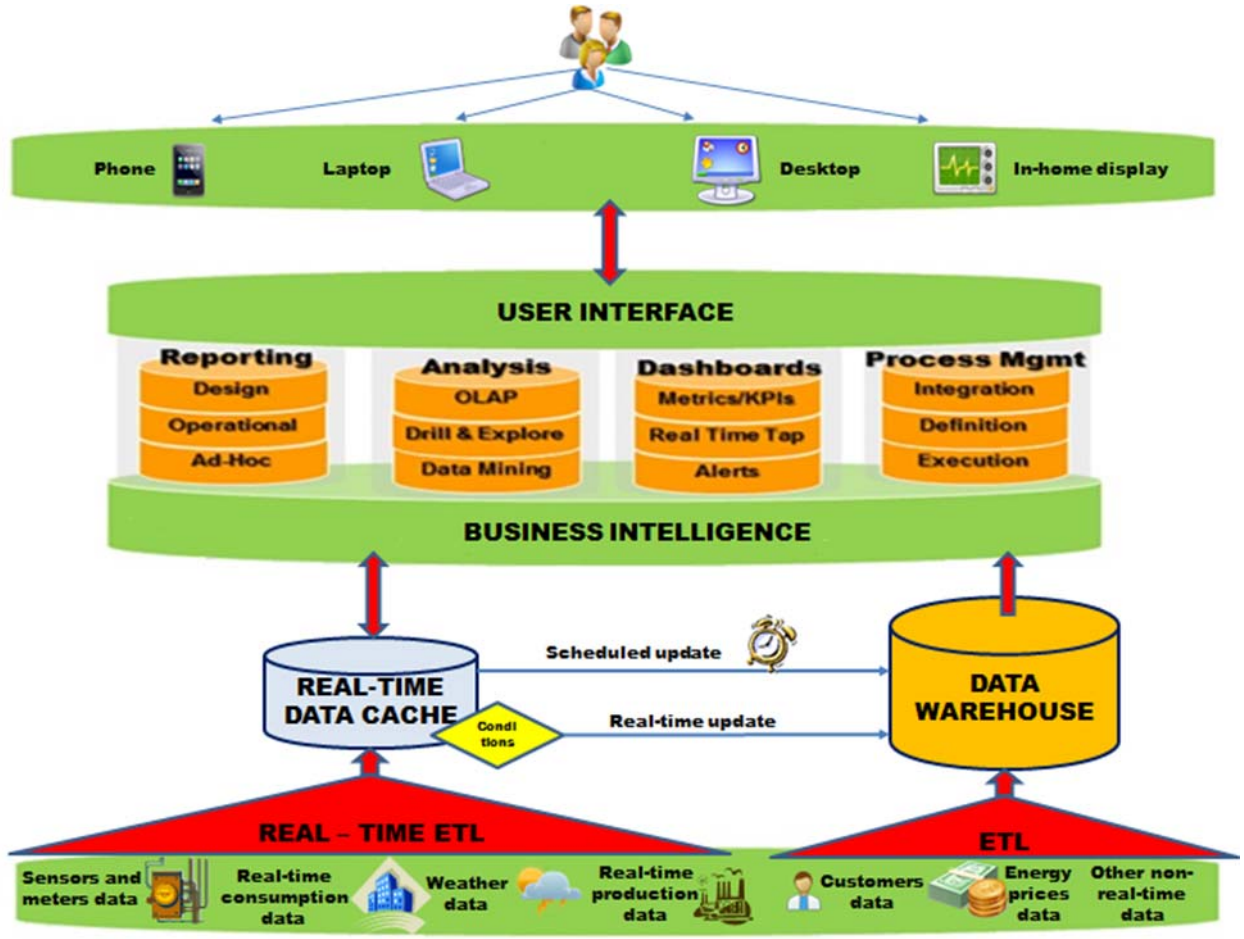


Fig.11. Our RTBI vision for the utilities industry

7. Conclusions

An ideal business intelligence system gives utilities the right information in the right format, at the right time, so they can make optimal business decisions. Giving the fact that this industry is an environment where decisions are time sensitive, utilities need RTBI solutions to improve customer experiences and operational efficiencies.

Our vision of real-time business intelligence for the utilities industry proposes the use of an external real-time data cache and a traditional data warehouse, in order to eliminate performance problems and to solve problems like internal inconsistency, scalability and data latency.

Although many technologies are available to implement real-time business intelligence, many challenges remain to make this a reality.

References

[1] Florian Neuhof, "Smart by necessity", 05.10.2010, http://www.utilities-me.com/article-817-smart-by-necessity/1/print/

[2] Purnendu Kumar, "Smart Grid…the next generation of electrical transmission and distribution system", 04.09.2012, http://purnendukumar.wordpress.com/2012/09/04/smart-grid-the-next-generation-of-electrical-transmission-and-distribution-system/

[3] ElegantJ BI, http://www.elegantjbi.com/solutions/industry_bi_utilities.htm

[4] Wayne W. Eckerson, Performance Dashboards**:** Measuring, Monitoring, and Managing Your Business, John Wiley & Sons, 07.10.2010 - 336 pp.

[5] Health Technology, http://www.healthtechnolgy.com/business-intelligence-systems.html

[6] Altek solutions,"Business Intelligence Solutions for Utilities and Energy", http://alteksolutions.com/solutions/industry/utilitiesenergy.asp

[7] Heena Gathibandhe et. al., "How Smart is Real-Time BI?", 28.01.2010, http://www.information-management.com/infodirect/2009_152/real_time_business_intelligence-10017057-1.html?zkPrintable=1&nopagination=1

[8] Wayne Eckerson, "The Soft Side of Real-Time BI", 01.08.2004, http://www.information-management.com/issues/20040801/1007215-1.html?zkPrintable=1&nopagination=1

[9] Daniela Ioana Sandu, "Operational and real-time Business Intelligence", Informatica Economică Journal 3(47)/2008, http://revistaie.ase.ro/content/47/06Sandu.pdf.

[10] B.S. Sahay, Jayanthi Ranjan, "Real time business intelligence in supply chain analytics", Information Management & Computer Security, Vol. 16 Iss: 1, pp.28 – 48, 2008.

[11] Colin White, "*Now is the Right Time for Real-Time BI"*, Information Management Magazine, September, 2004.

[12] Judith R. Davis, "Right-Time Business Intelligence: Optimizing the Business Decision Cycle", Sybase, 2006

[13] B Azvine, Z Cui, D D Nauck, B Majeed, "Real Time Business Intelligence for the Adaptive Enterprise," cec-eee, pp.29, The 8th IEEE International Conference on E-Commerce Technology and The 3rd IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services (CEC/EEE'06), 2006, http://bingo.crema.unimi.it/turing/MATERIALE/admin/allegati/architetturaBT.pdf

[14] Wikipedia, "Real-time business intelligence", http://en.wikipedia.org/wiki/Real-time_business_intelligence

[15] EBN, "IDC: Utilities Need to Implement Sophisticated Business Intelligence Tools",

EBN, 13.04.2011, http://www.ebnonline.com/document.asp?doc_id=205629
[16] Justin Kern, "Utilities Surge Ahead with BI, Analytics Despite Immaturity and Integration Problems", 20.09.2012, http://www.information-management.com/news/utilities-surge-ahead-with-BI-analytics-despite-immaturity-problems-10023203-1.html
[17] Timo Elliott, "Turn Big Data indo Business Value with Real-Time BI", March 2012, http://assets.timoelliott.com/docs/innovationbysapbigdata_nordic.pdf
[18] Azvine, B., Cui, Z. and Nauck, D. (2005), "Towards real-time business intelligence", BT Technology Journal, Vol. 23 No. 3, pp. 214-25
[19] Judith. R. Davis „Right Time Business Intelligence
Optimizing the Business Decision Cycle", research report published by
BEYE Research, December 2005.
[20] Oracle Utilities Business Intelligence – user guide, http://docs.oracle.com/cd/E23251_01/pdf/E18759_04.pdf

Janina POPEANGĂ graduated in 2010 from the Faculty of Cybernetics, Statistics and Economic Informatics, Economic Informatics specialization, within Academy of Economic Studies of Bucharest. The title of her Bachelor's thesis is "Distributed databases". In 2012, she graduated the Databases for Business Support master program with a thesis entitled "Monitoring and management of electric power consumption using sensorial data". Janina's interests are broadly in the fields of databases and distributed systems. She is now planning to begin her PhD, advised by Professor Ion LUNGU. Her research focuses on real-time database systems, business intelligence analytics, sensor data management, smart grid and renewable energy.

A Framework for Automated Database TuningUsing Dynamic SGA Parameters and Basic Operating System Utilities

Hitesh KUMAR SHARMA1, Aditya SHASTRI2, Ranjit BISWAS3
1Assistant Professor, University of Petroleum & Energy Studies
2Vice-Chancellor, Banasthali University, Rajasthan-304022, India
3Head and Professor,CSE Dept, Jamia Hamdard (Hamdard University)
hkshitesh@gmail.com, adityashastri@yahoo.com, ranjitbiswas@yahoo.com

In present scenario the manual work (Done by Human) cost more to an organization than the automatic work (Done by Machine)and the ratio is increasing day by day as per the tremendous increment in Machine (Hardware + Software) Intelligence. We are moving towards the world where the Machines will be able to perform better than today by their own intelligence. They will adjust themselves as per the customer's performance need. But to make this dream true, lots of human efforts (Theoretical and Practical) are needed to increase the capability of Machines to take their own decision and make the future free from manual work and reduce the working cost. Our life is covered with the different types of systems working around. The information system is one of them. All businesses are having the base by this system. So there is the most preference job of the IT researcher to make the Information system self-Manageable. The Development of well-established frameworks are needed to made them Auto-tuned is the basic need of the current business.The DBMS vendors are also providing the Auto-Tune packages with their DBMS Application. But they charge for these Auto-Tune packages. This extra cost of packages can be eliminated by using some basic Operating system utilities (e.g. VB Script, Task Scheduler, Batch Files, and Graphical Utility etc.).

We have designed a working framework for Automatic Tuning of DBMS by using the Basic Utilities of Operating System (e.g. Windows) .These utilities will collect the statistics of SGA dynamic Parameters. The Framework will automatically analyze these SGA Parameter statistics and give suggestions fordiagnose the problem. In this paper we have presented that framework with practical Implementation.

Keywords*: SGA, SGA Dynamic Parameters, Database Tuning, DBA, Automated Tuning, TOC.*

1 Introduction

As we have seen, hardware costs fall rapidly while human costs remain relatively static. This leads to a condition there the human costs of manual tuning activities outpaces the costs of faster hardware .
(seeFigure 1).

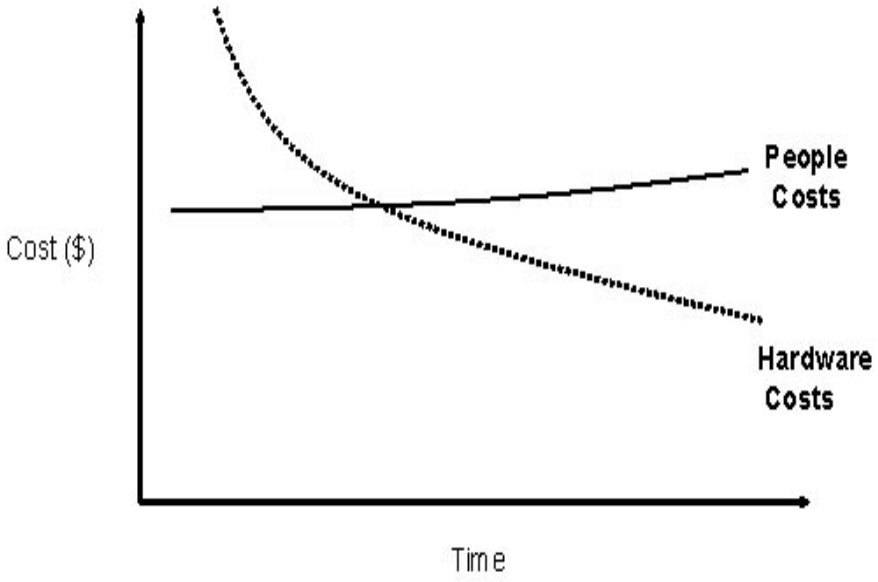


Fig 1: H/W cost vs Human Cost [Ref. 5]

Most large databases are managed by DBAs who are responsible for the good performance of the database but manual physical design is both time consuming and very tedious, as the database administrator (DBA) needs to find the benefits of different individual design features that can possible interact with one another. Motivated not only by the difficulty of tuning but also from the need to reduce the total cost of ownership in their products, several commercial DBMS vendors offer automated tools with several features but the cost for ownership of these tools is also high.With the dramatic drop of hardware and software prices, the expenses due to human administration and tuning staff dominate the cost of ownership for a database system [1].

2. Manual Tuning Framework

Database Administrator is responsible for enhancing the performance of database system. The detection of performance degradation is achieved by continuously monitoring system performance parameters. Several methods including the usage of materialized views and indexes, pruning table and column sets, usage of self healing [2] Techniques, usage of physical design tuning etc. have been proposed that proactively monitor the system performance indicators, analyze the symptoms and auto tune the DBMS to deliver enhanced performance. The performance degradation is due to increased workload on the system. This increased load has to be minimized to enhance the response rate of the system. In order to achieve this objective, either the administrator decreases some amount of load by closing some files or he may increase the RAM. The administrator has to check continuously or we can say, at regular intervals the Buffer Cache Hit Ratio(BCHR) [4]. Based on this hit ratio, the database administrator determines if more amount of

RAM has to be allocated. This task of load reduction by increasing RAM requires manual intervention and thus may take even years to complete [2].

(see figure 2)

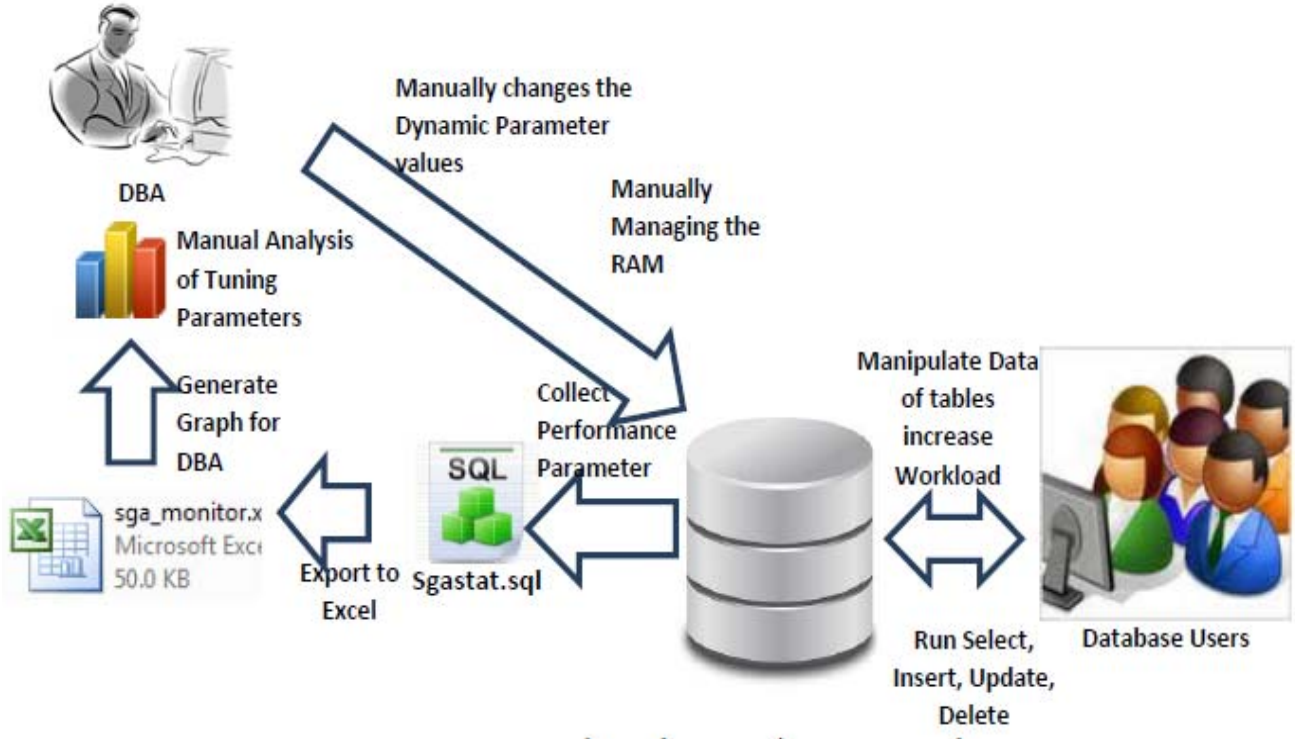


Fig 2: Manual Database Tuning Framework

However, Oracle manages RAM memory demands according to the demands of each task by using sophisticated algorithms to improve the speed of RAM intensive tasks. Oracle DBA can dynamically de-allocate RAM memory as well as re-allocate it. But since database administrator is a normal human being, he cannot calculate the actual amount of RAM memory required by an application. [5]

Due to this limitation of DBA, the allocation of RAM manually for optimizing performance of database system becomes a complicated as well as costly task.

3. Building Blocks of Automation Tuning Framework

Many business applications demand the use of complex database systems which should be administered and optimized for better performance. As suggested in [2], physical tuning should be avoided as it is expensive. As the physical design of database suffers from various limitations, an automated database tuning framework is proposed in order to achieve high grade of performance. The Framework is employed for identifying the symptoms and altering key system parameters.

The Framework has three basic building:(Complete Frameworkhas shown in figure6)

a. Automated Workload Generation Block
b. Working Database

c. Automated Database Tuning Block

This block will be used for generation of variable workload on working database by creating virtual users.

3.1. **Automation Workload Generation Block**

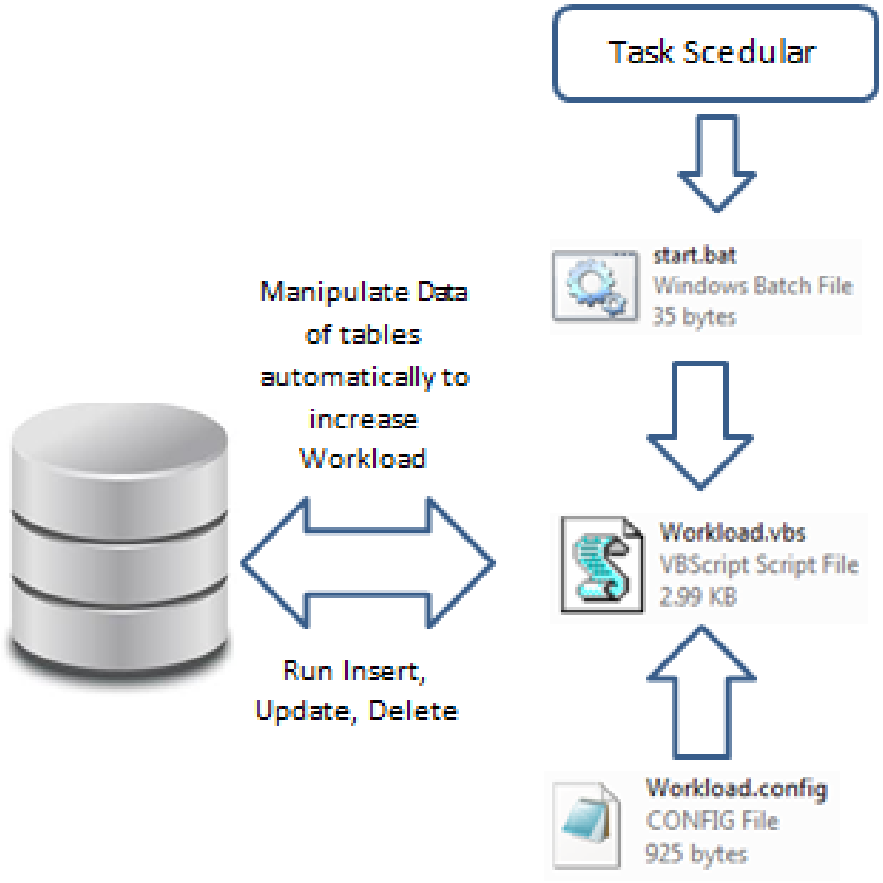


Fig 3: Automatic Workload Generation block

This block will use basic utilities of operating system (i.e. Vb Scripting, Config file etc.) to create virtual users and provide login accessibility to those virtual user for database.(See figure 3). The VB Script will increase the workload on working database as per the setting written in Workload.config file. By using these two files we can create virtual workload on working database and simultaneously achieve the automated generation of variable workload.

3.2. Working Database

This block is the working database on which other two blocks will perform their respective functions automatically (see figure 4):

Fig 4: Working Database

The working database is represented by the general symbol used for any Database.

3.3. Automation Database Tuning Block

This block is the soul of the complete framework. This block will collect the statistic of SGA Dynamic parameters automatically and insert to a monitoring table. On the basis of this table an automation application will correlate the parameters value to the performance of the Database.

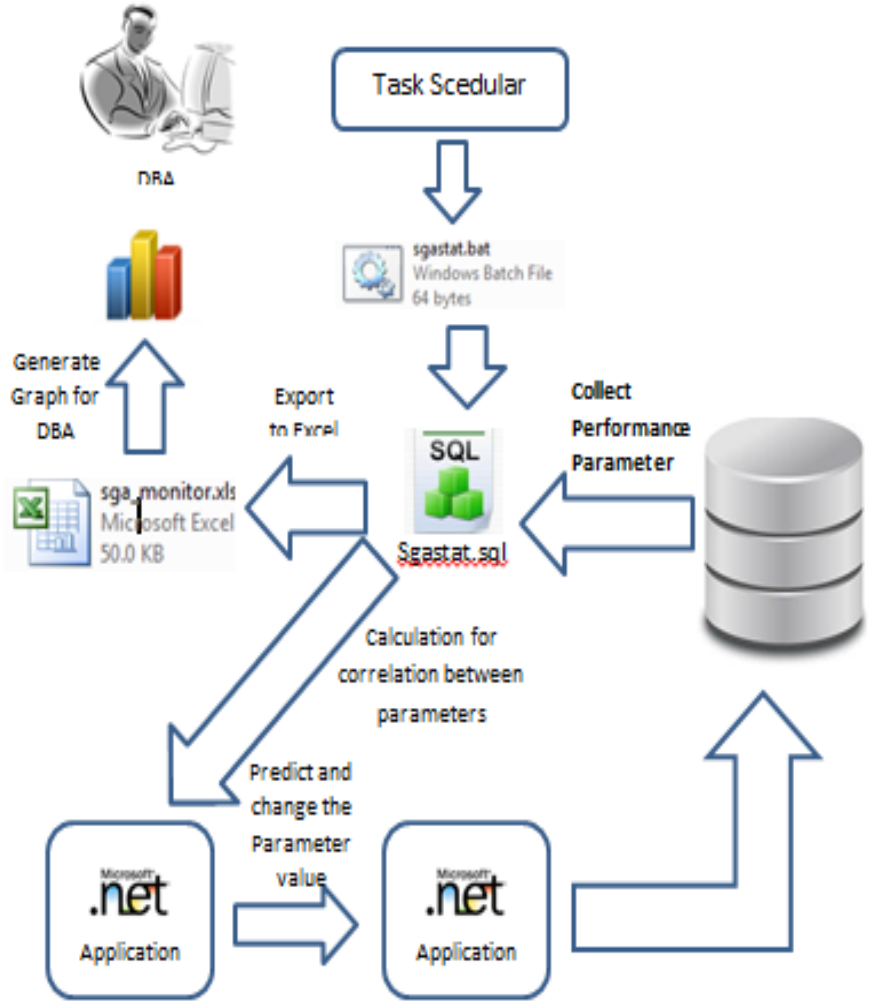


Fig 5: Automatic Tuning Block

VB script through the Schedular. The SQL script will fetch the SGA parameters statistics and insert in sga_monitor table.

After analysis of the parameters value the automation application will automatically decide to change the value of only those parameters whose changed value will increase the performance of the database. (See figure 5)
This table data will be used by a .Net application for analysis. After analyzing the values the same .net application will take self-decision to change the value of SGA parameters. One part of this block will also create the charts by using excel utility for DBA (Optional Block). This will help to DBA to find the automatic tuning is doing the job effectively or not.
These blocks will tune the database using various tuning rules as well as system parameters. However, several parameters can be altered simultaneously for better performance gain. The third block estimates the required value of dynamic SGA parameter based on the current DBMS input parameters and applies the necessary correction to change the size of these parameters based on the tuning rules.

4. Complete Framework of Automation Tuning

The Complete framework has shown in figure 6this framework is the combination of the above defined three blocks. After implementing this framework there will be no need for manual tuning. The combination of Basic Utilities of Operating System and database statistics an organization can achieve the automation in Database tuning.

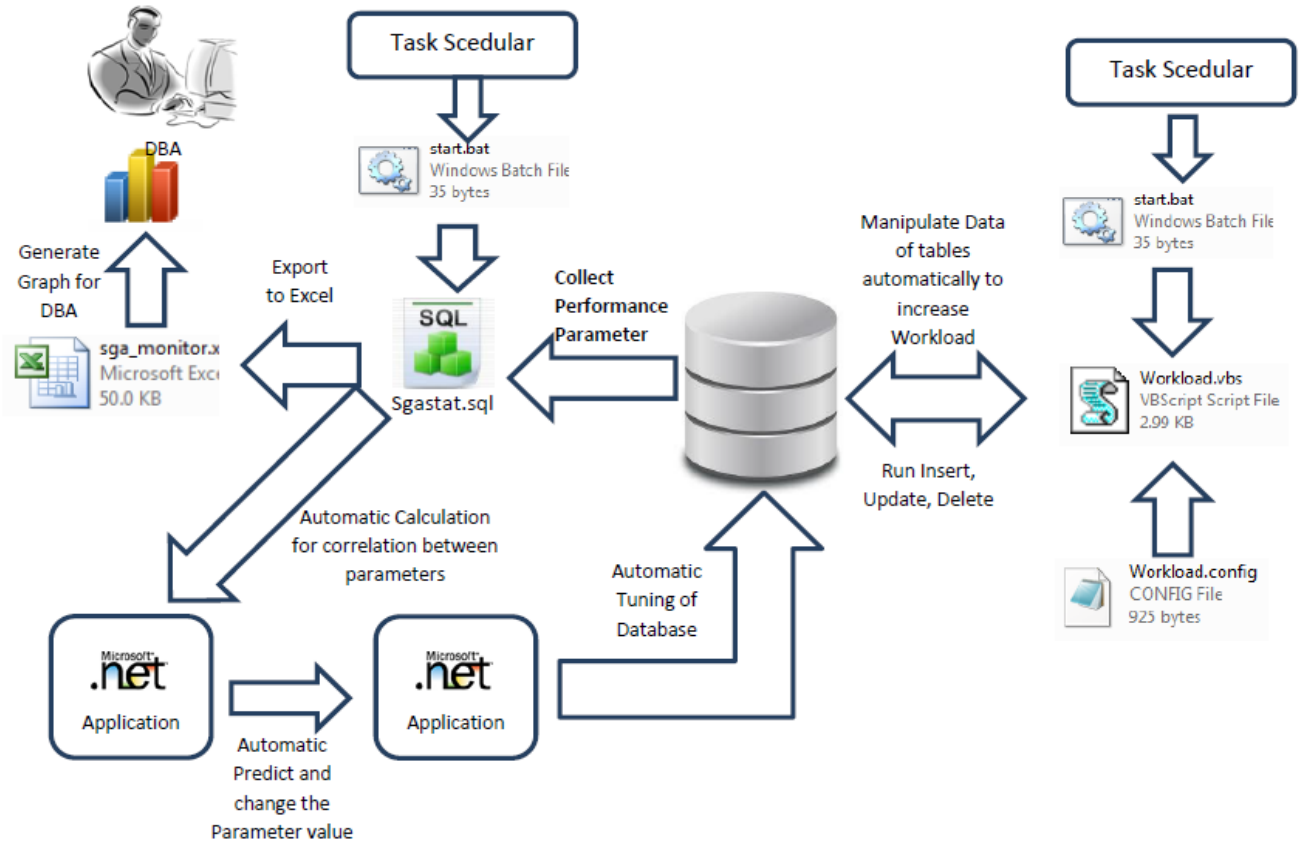


Fig 6: AutomatedDatabase Tuning Framework

The important thing about this framework is that it will be faster than other tuning utilities due to usage of Operating System basic utilities. These basic utilities are directly connected to the kernel of OS. So it will give faster result than other framework.

5. Conclusion and Future Work

Tuning the database can become quite complex, but modern databases offers the administrator an unparalleled ability to control the PGA and SGA. Until old databases evolves into a completely self-tuning architecture, the DBA was responsible for adjusting the dynamic configuration of the system RAM. Automated SGA adjustment scripts can be used to allow the DBA to grow and shrink the SGA regions. Manual tuning cost more for an organization but it is one of the major need for an organization to attract the customer. So we have proposed a solution to fulfill the need of an organization in the shape of this Automation Framework. This framework will not take any cost and it will give faster result compare to manual tuning.

The future work is to implement this framework and test in a working environment. The work is in progress and we will come with the result very soon in future.

References

[1] Sharma H., Shastri A., Biswas R. **"**Architecture of Automated Database Tuning Using SGA Parameters", Database System Journal, Romania, 2012

[2] K. P. Brown, M. Metha, M. J. Carey, and M. Livny, "Towards Automated Performance Tuning for Complex Workloads", Proceedings of 20th VLDB Conference, Santiago, 1994

[3] G. Weikum, A. C. Konig, A. Kraiss, and M. Sinnewell, "Towards Self-Tuning Memory Management for Data Servers", In Bulletin of the Technical Committee on Data Engineering, IEEE Computer Society, Vol. 22, No. 1, 1999.

[4] T. Morals and D. Lorentz, "Oracle 9i: Database Reference", Oracle Corporation, 2001.

[5] Donal k. Burleson, "The Definitive Reference", Second Edition.

About Authors:

Hitesh KUMAR SHARMA, The author is an Assistant Professor in University of Petroleum & Energy Studies, Dehradun. He has published 8 research papers in National Journals and 3 research paper in International Journal. Currently He is pursuing his Ph.D. in the area of database tuning**.**

Aditya SHASTRI, Ph.D. MIT, Published about 200 research papers in international journals on Graph Theory with applications in Communication, Computer Graphics and Parallel Processing ,Vice Chancellor, Director, BanasthaliUniversity, Banasthali, INDIA

Ranjit BISWAS, Head and Professor Jamia Hamdard (Hamdard University) Published about 100 research papers in International journals/bulletins.

A Multidimensional View Proposal of the Data Collected Through a Questionnaire. Associated Data Mart Deployment Framework

Mihaela I. MUNTEAN, Diana TÂRNĂVEANU
Department of Business Information Systems
West University of Timişoara, Faculty of Business Administration
Timişoara, ROMÂNIA
mihaela.muntean@feaa.uvt.ro, diana.tarnaveanu@feaa.uvt.ro

Beyond the traditional data analysis approaches based on SPSS (or similar statistical software tools), an alternative demarche will be subject of our debate. Performant data analysis can be completed based on a multidimensional view of the collected data. This implies an additional data mart powered with information obtained through an ETL process from the collected data. Measures and dimensions will facilitate a subject-oriented, time-based analysis. The theoretical approach framework for deploying the data mart will ground a multidimensional analysis on „how the different respondents answered to the questions included into the questionnaire?". In addition, a study case was proposed, a questionnaire built and different analyses presented.

Keywords: *multidimensional model, questionnaire, data mart, ETL, data analysis*

1 Introduction

D. Slesinger and M. Stephenson in the Encyclopedia of Social Sciences define research as "the manipulation of things, concepts or symbols for the purpose of generalizing to extend, correct or verify knowledge, whether that knowledge aids in construction of theory or in the practice of an art."

A research process consists of a number of closely related activities like: (1) formulating the research problem; (2) extensive literature survey; (3) developing the hypothesis; (4) preparing the research design; (5) determining sample design; (6) **collecting the data**; (7) execution of the project; (8) **analysis of data**; (9) hypothesis testing; (10) generalizations and interpretation, and (11) preparation of the report or presentation of the results, i.e., formal write-up of conclusions reached.

Regarding the collection of data, the researcher should select the proper method of collecting the data taking into consideration the nature of investigation, objective and scope of the inquiry, financial resources, available time and the desired degree of accuracy [1]. **Data can be collected:** (a) by observation; (b) through personal interview; (c) through telephone interviews; (d) **by mailing of questionnaires; using another way of distribution; handling web-based questionnaires**; (e) through schedules [2].

A questionnaire is a set of carefully designed questions given in exactly the same form to a group of people in order to collect data about some topics in which the researcher is interested [3]. As with any other branch of science, the validity and reliability of the measurement tool, i.e. the questionnaire, needs to be rigorously tested to ensure that the data collected is meaningful. The design and the administration method of a questionnaire will also influence the response rate that is achieved and the quality of data that is collected. A questionnaire is given out (normally quantitative) to gather statistical data about responses; in some situations it is recommended to do research in more depth by interviewing (normally qualitative) selected members of the questionnaire sample.

Nowadays, web-based questionnaires are a fiable solution, suitable to distribute, and to collect the opinions of a large number of respondents. The data collected is carefully

analyzed and decisions are grounded. Statistical questionnaires serve to understand patterns and to identify trends within the data. Also, by analyzing the data, predictions should be enabled.

Beyond the traditional analysis approaches based on SPSS (or similar statistical software tools), an alternative demarche will be subject of our debate. **Performant data analysis can be completed based on a multidimensional view of the collected data** [4]. This implies **an additional data mart** (Figure 1) powered with information obtained through an ETL (Extract-Transform-Load) process from the collected data. Measures and dimensions will facilitate a subject-oriented, time-based analysis.

Fig. 1. The data mart environment

2. Approaching questionnaires

2.1 Theoretical background

Looking forward to the desired information and having in mind how data will be collected, questions will be established. Questions can be divided into those **directly related to the research question** (research subject); **filter questions** that explore the characteristics of the different study groups and **'filler' questions** that, although not part of the research question, aid the flow of the questionnaire. Questions can be formulated as **open-ended questions**, where the respondent is free to give her/his own response to the questions, or **closed questions**, where a choice of predetermined answers is given. In both cases, the wording of the questions has an important influence on the responses that are given.

The **layout of the questionnaire** is important not only for ensuring that all the questions are answered, but also for facilitating data coding and analysis. It's important that people's attention is captured and that they are interested in completing the questionnaire.

It is very important to **pre-pilot the questionnaire** for identifying any ambiguities of the questions and the range of possible responses for each question. It is not a formal procedure, more an information-gathering exercise [5]. The next step is to perform a **pilot study to test the questionnaire**, this phase being important for verifying if the data collected from the questionnaire is valid and reliable. At last, but not at least, the **coding schema for questions and answers** will be established. It has a major impact on the ETL process.

2.2 Questionnaire proposal

The Faculty of Economics and Business Administration (FEAA) has a new website – www.feaa.uvt.ro. We share information with a wide audience, including academics, researchers, current and prospective students, parents, business environment and media representatives, and many other visitors. Their feedback is important and may conduct to improvements or even to a reengineering of the whole, but with respect to the „spirit of FEAA".

The proposed online questionnaire (Figure 2) basically **manages to collect two categories of data**: (1) feedback regarding the quality of the website, and (2) **consumer demographic data.**

Generally, on-line questionnaires are used by businesses across the globe to survey potential customers, business partners and recent clients. The gathered responses and data interpretation can be invaluable to a marketing campaign or business venture. When users create online questionnaires, they have a number of survey questionnaire templates that they can choose from that will help them launch a professional survey without delay. There are many factors in designing an online questionnaire; guidelines, available question formats,

administration, quality and ethic issues should be reviewed [6].

Our proposed online questionnaire includes mainly closed questions related to: (1) the general impression, 2) the aspect, (3) the content, (4) the layout, (5) ease of navigation, (6) the necessity of displaying commercial advertisements, (7) problems when downloading files, (8) the number of pages accessed, (9) the time spent on the website, (10) if the respondent expectancies were fulfilled, (11) the frequency of accessing the website, (12) the reason for the first visit, (13) either or not a subscription to our University's Press Magazine is wanted, (14) if the subject is a FEAA alumni, (15) if the subject wants to be a part of the FEAA community, (16) the necessity of displaying available jobs, (17) either or not the parents should have access to the students records, (18) a self-evaluation of the subject's skills on browsing the web, (19) age, (20) position, (21) how much time the subject was employed, (22) city, (23) the preferred web browser, (24) the operating system, (25) citizenship and (26) marital status. The date and time when the respondent completed the questionnaire are also recorded.

Three more open-ended questions were added: the e-mail in case the subject wants to receive the results of the study, proposed changes to the website, and what other information/characteristics would she/he like to see on the website.

The responses to the last two questions were recorded and sent to the website developer, but ignored in the study conducted further.

Fig. 2. The proposed questionnaire

A number of 131 responses were recorded, users who accessed the site during two school weeks, taking into account that the questionnaire wasn't mandatory.

3. Multidimensional view of the data collected through a questionnaire

3.1 Data mart/warehouse general considerations

According to McKnight W., establishing the proper data warehouse or data mart architecture is quite challenging [7]. The efficacy of having a centralized data store with quality, integrated, accessible, high performance and scalable data can't be denied, but short term business needs and other interests can conduct to a data mart oriented approach.

The data warehouse/data mart must enclose items/objects of importance to the business as customer, product, time, geography, sales hierarchy and market (referred to as

'dimensions' since they define the context of the business transactions).

Practically, the data warehouse/data mart is a database in which atomic level data from disparate sources is brought together in a structured way creating one multi-subject oriented version of the corporate/department truth, designed to enable timely, accurate decision making in support of strategic and tactical business initiatives.

A **star, snowflake** or **constellation schema**, consisting of facts and dimension tables, will be generated for the data warehouse/data mart, grounding the multidimensional cube deployment process [8].

Having a look at the star schema in Figure 3, the following considerations should be kept in mind:

(1) records in the fact table are often referred to as events, due to the time-variant nature of a data warehouse/data mart environment;
(2) the fact table is linked to all surrounding dimensions;
(3) the primary key of the fact table is defined by the set of foreign keys introduced to join the fact table with all independent dimensions – D_1, D_2, ……, D_n;
(4) the fact table contains the measures of the analysis - M_1, M_2,…., M_p, where $M_k = f(D_1, D_2,…, D_n)$, having k = 1, 2, ….., p;
(5) dimensional attributes are added to describe dimensional values.

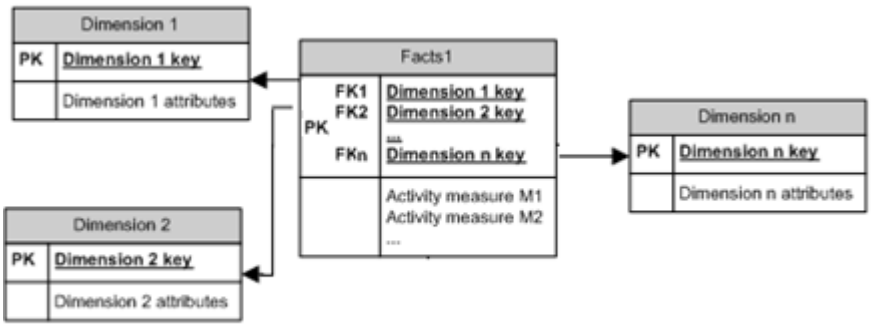


Fig. 3. Data warehouse/data mart star schema

The granularity or data grain of a fact is the level of detail at which the respective fact is recorded (i.e. the level of detail for the measurement) and made available to the dimensional model. Granularity represents a very important aspect in the analysis of data, as it determines the level of available information. For this reason, all recorded data should be kept at a highest granularity level (i.e. highest level of detail) that may be easily changed into a lower level through summarization.

Hierarchies represent the base structures of dimensions and define the relationship between the existing attributes of different levels (Figure 4). They are used for analytical processing in the data warehouse/data mart environment and visualization of data at different aggregation levels.

Dimensional hierarchies will enrich the model and will enable roll-up and drill-down operations on the multi-dimensional cube. The drill-down operation refers to moving downwards along the hierarchical levels of a dimension in order to obtain more details at a lower level of granularity. By contrast, the roll-up operation refers to moving upwards along the hierarchical levels of a dimension in order to achieve less details or a higher level of granularity (i.e. aggregated data).

One dimension can host different hierarchies, all of them starting from the same base level.

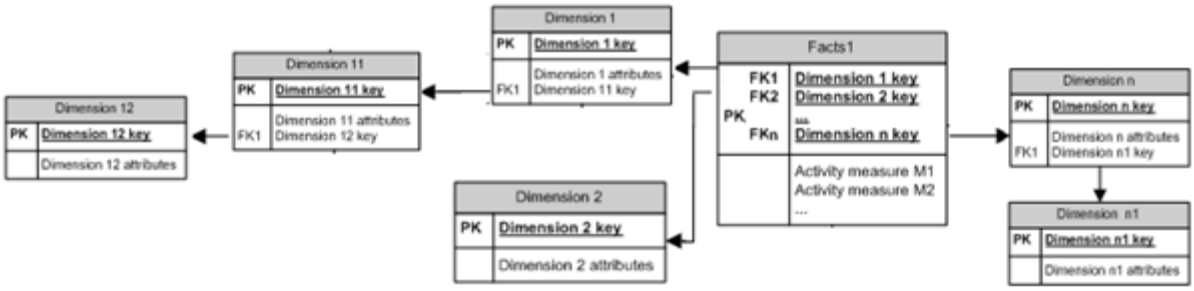


Fig. 4. Data Warehouse/data mart snowflake schema

A complex analysis requires a variety of measures which are depending on different dimensions (Figure 5). These measures are placed in distinct fact tables and are surrounded by the corresponding dimensions.

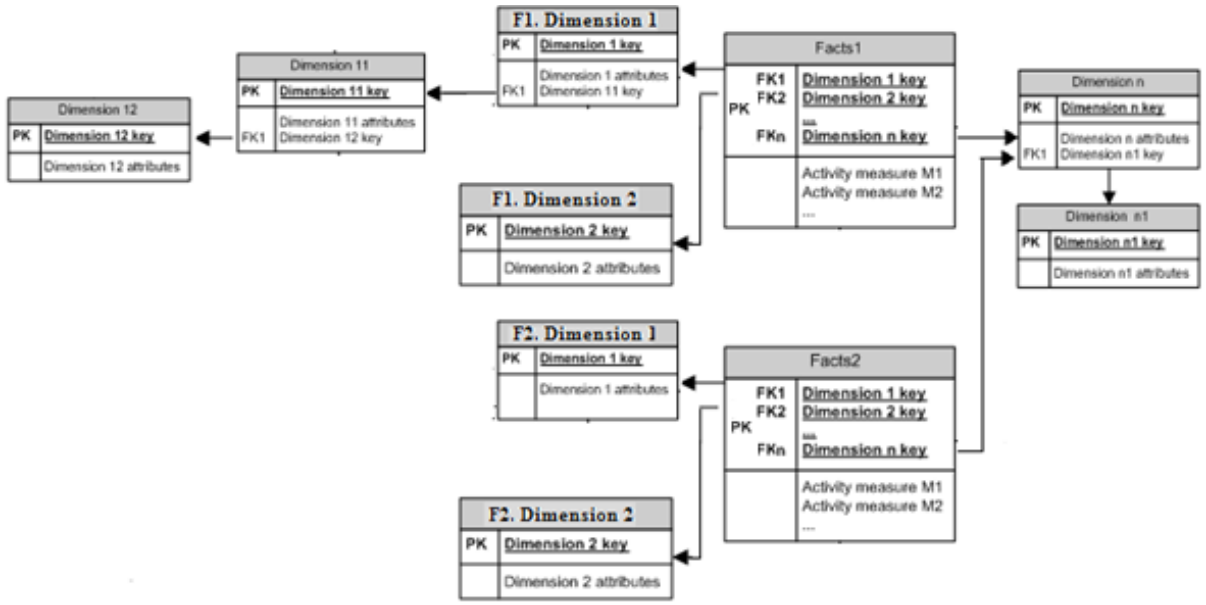


Fig. 5. Data warehouse/data mart constellation schema

There is no direct connection between the facts tables, associations can be made through a common dimension.

3.2 Data Mart Deployment Framework

A general approach framework for the introduced data mart environment (see Figure 1) will be proposed. Conveniently, under **source systems** we understand the database which will store all the answers of the respondents to the different questions ('**n**' questions directly related to the research subject and '**p**' filter questions) included in the questionnaire. Tables like QUESTIONS, VARIANTS, RESPONDENTS and ANSWERS could be part of the database. The following **database schema** is acceptable (Figure 6).

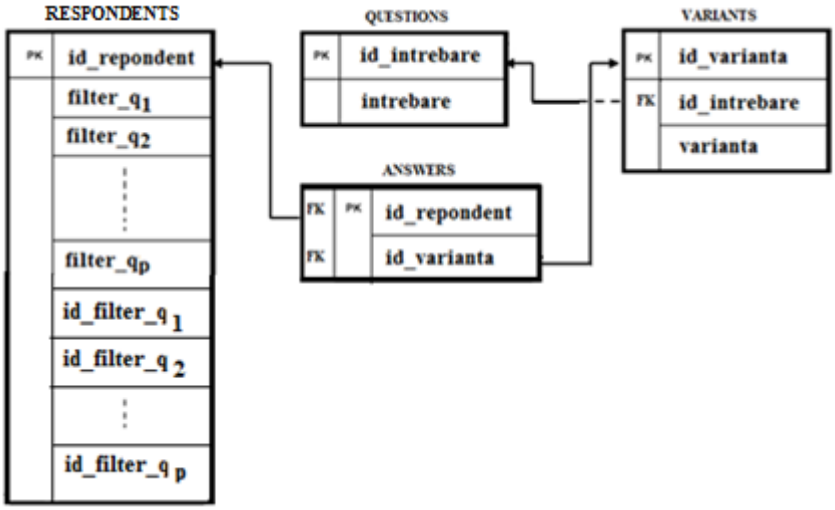


Fig. 6. Source database [4]

(1) **Questions directly related to the research** subject are usually formulated as closed questions and are stored in table QUESTIONS. Answers alternatives are stored in table VARIANTS.

(2) **Filter questions, closed or open-ended**, are exploring the characteristics of the respondents, and therefore will establish the alphanumeric fields filter_q_1, filter_q_2, …, filter_q_p of table RESPONDENTS.
The additionally fields id_filter_q_1, id_filter_q_2, …, id_filter_q_p will facilitate the developing of the data mart (Figure 9) and will be updated during the implementation of the ETL process.

(3) For each filled questionnaire, a new record will be added into table RESPONDENTS and '**n**' records will be added into table ANSWERS.

```
CREATE TABLE QUESTIONS (id_intrebare BYTE PRIMARY KEY, intrebare VARCHAR2 (25));

CREATE TABLE VARIANTS (id_varianta BYTE PRIMARY KEY, id_intrebare BYTE REFERENCES
       questions (id_intrebare), varianta VARCHAR2 (15));

CREATE TABLE RESPONDENTS (id_repondent INTEGER PRIMARY KEY,
       filter_q1 VARCHAR2 (…), filter_q2 VARCHAR2 (…), …, filter_qp VARCHAR2 (…),
       id_filter_ q1 BYTE, id_filter_ q2 BYTE, …, id_filter_ qp  BYTE);

CREATE TABLE ANSWERS (id_repondent INTEGER REFERENCES respondents (id_repondent),
       id_varianta BYTE REFERENCES VARIANTS (id_varianta),
       PRIMARY KEY (id_repondent, id_varianta));
```

According to [4], the data mart will ground a multidimensional analysis on „**how the different respondents answered to the questions included into the questionnaire?**" Developing a data warehouse/data mart is quite challenging, several development methodologies have been identified [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21].
The design of a representative dimensional model (i.e. data mart) can be performed within an agile framework (adapted from [22]) (Figure 7).

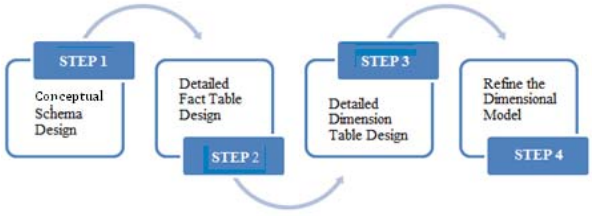


Fig. 7. Data mart agile development framework

STEP1. Conceptual schema design

(1) The variants of the '**n**' questions that are directly related to the research subject will represent dimensional values; i.e. the variants v_{k1}, v_{k2},…, v_{km} of question '**k**' will represent dimensional values on the 'k' dimension - D_k, k = $\overline{1,n}$.

(2) Consumer demographic data (from the filter questions) will add '**p**' new dimensions into the model d_{q1}, d_{q2},…, d_{qp}.

(3) For each question we will have a distinguished measure (**M_1, M_2, …, M_k,…, M_n**), which will be aggregated taking into consideration the corresponding variants dimension and all dimensions introduced by the filter questions (Figure 8).

$$\mathbf{M_k = f\,(D_k, d_{q1}, d_{q2}, \ldots, d_{qp}),}\ \ k=\overline{1,n}.$$

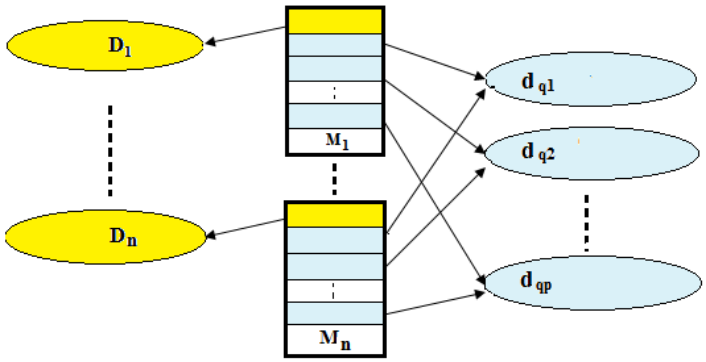


Fig. 8. Data mart conceptual schema

STEP2. Detailed fact table design

The first step in the design of the dimensional model's fact table consists of determining its key. A defining characteristic of the dimensional model is that the fact table has a non-minimal composite key, comprising all dimension tables' primary keys. Aside from its key, the fact table contains measures or facts that may be analyzed from the various perspectives described by its surrounding dimensions (Figure 9).

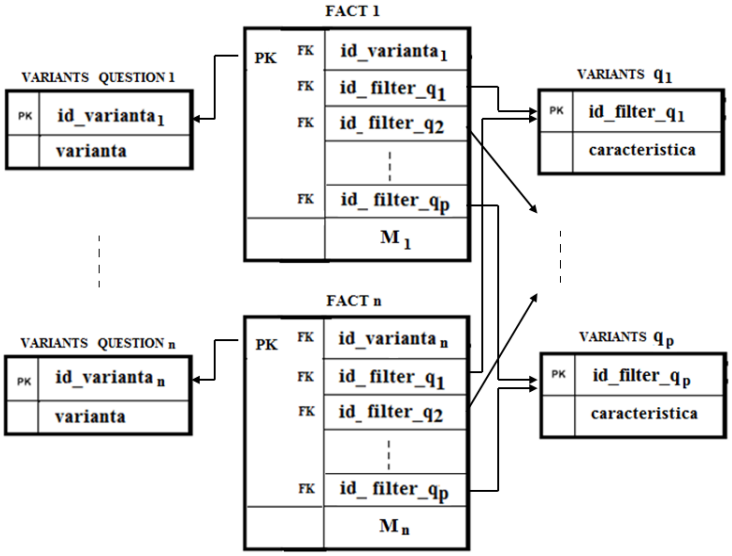


Fig. 9. Detailed data mart schema [4]

STEP3. Detailed dimension table design

An essential aspect in the detailed dimension table design step, and generally in

dimensional modeling, is the identification and representation of hierarchies, which define the basis of the aggregation and the analysis processes.

STEP4. Refine the dimensional model

The dimensional modeling activity has to be accompanied by careful assessment of the end-user informational needs and the underlying data supply. Given these arguments, the model will be refined in order to enhance the analysis possibilities and provide a better and simpler model for the decision makers to use.

The **effective deployment of the designed data mart** can be realized with the help of some proper Oracle SQL scripts:

(1) one script for creating all the implied tables;
(2) another script (s) for implementing the ETL process.

SCRIPT 1

```
CREATE TABLE variants_question_1(id_varianta_1 BYTE PRIMARY KEY,
       varianta VARCHAR2(…));
…
CREATE TABLE variants_question_n(id_varianta_n BYTE PRIMARY KEY,
       varianta VARCHAR2(…));

CREATE TABLE variants_q1(id_filter_q1 BYTE PRIMARY KEY,caracteristica VARCHAR2(…));
…
CREATE TABLE variants_qp(id_filter_qp BYTE PRIMARY KEY,caracteristica VARCHAR2(…));

CREATE TABLE fact1(id_varianta_1 BYTE REFERENCES variants_question_1
       (id_varianta_1), id_filter_q1 BYTE REFERENCES variants_q1(id_filter_q1),…,
       id_filter_qp BYTE REFERENCES variants_qp (id_filter_qp),
       M1 INTEGER, PRIMARY KEY (id_varianta_1, id_filter_q1,…,id_filter_qp));
…
CREATE TABLE factn (id_varianta_n BYTE REFERENCES variants_question_n
       (id_varianta_n), id_filter_q1 BYTE REFERENCES variants_q1(id_filter_q1),…,
       id_filter_qp BYTE REFERENCES variants_qp (id_filter_qp),
       Mn INTEGER, PRIMARY KEY (id_varianta_n, id_filter_q1,…,id_filter_qp));
```

SCRIPT 2

```
INSERT INTO variants_question_1  SELECT  id_varianta, varianta FROM variants
       WHERE id_intrebare = 1;
…
INSERT INTO variants_question_n  SELECT id_varianta, varianta FROM variants
       WHERE id_intrebare = <value_of_n>;

DECLARE
/* dimension VARIANTS q1 - generating dimensional values */
/* updating field id_filter_q1 in table RESPONDENTS      */
v_q1               BYTE;
v_q1_varianta      VARCHAR2(…);
CURSOR c  IS  SELECT  filter_q1 FROM respondents GROUP BY filter_q1;
BEGIN
       SELECT COUNT (DISTINCT filter_q1) INTO v_q1 FROM respondents;
       OPEN c;
       FOR  i  IN   1.. v_q1 LOOP
              FETCH c INTO v_q1_varianta;
              INSERT  INTO variants_q1  VALUES (i, v_q1_varianta);
              UPDATE respondents SET id_filter_q1 = i  WHERE filter_q1 =
              v_q1_varianta;
       END LOOP;
       CLOSE;
END;
…
DECLARE
/* dimension VARIANTS qp -  generating dimensional values   */
/* updating field id_filter_qp in table RESPONDENTS         */
v_qp               BYTE;
v_qp_varianta      VARCHAR2(…);
```

```
CURSOR c  IS  SELECT  filter_qp FROM respondents GROUP BY filter_qp;
BEGIN
       SELECT COUNT (DISTINCT filter_qp) INTO v_qp FROM respondents;
       OPEN c;
       FOR  i  IN   1.. v_qp LOOP
              FETCH c INTO v_qp_varianta;
              INSERT  INTO variants_qp  VALUES (i, v_qp_varianta);
              UPDATE  respondents  SET id_filter_qp =  i  WHERE filter_qp =
              v_qp_varianta;
       END LOOP;
       CLOSE;
END;

/*  determining measure M1    */
SELECT  variants.id_varianta, id_filter_q1,…,id_filter_qp, COUNT(id_repondent)
       FROM  respondents, variants, answers
       WHERE respondents.id_repondent  =  answers.id_repondent AND
              answers.id_varianta  =  variants.id_varianta AND id_intrebare = 1
      GROUP BY variants.id_varianta,id_filter_q1,…,id_filter_qp;
…
/* determining measure  Mn    */
SELECT  variants.id_varianta, id_filter_q1,…,id_filter_qp, COUNT(id_repondent)
       FROM  respondents, variants, answers
       WHERE respondents.id_repondent  =  answers.id_repondent AND
              answers.id_varianta  =  variants.id_varianta AND
              id_intrebare  =  <value_of_n>
       GROUP BY variants.id_varianta,id_filter_q1,…,id_filter_qp;
```

3.3 Analyzing data collected through the proposed questionnaire

The theoretical approach presented in paragraph 3.2 was applied in order to analyze the data collected through the questionnaire introduced in Figure 2, representing a future step of the researched introduced in [23].

The snowflake schema for analyzing the first question is chosen for detailed exemplification (Figure 10), corresponding information is displayed in Figure 11.

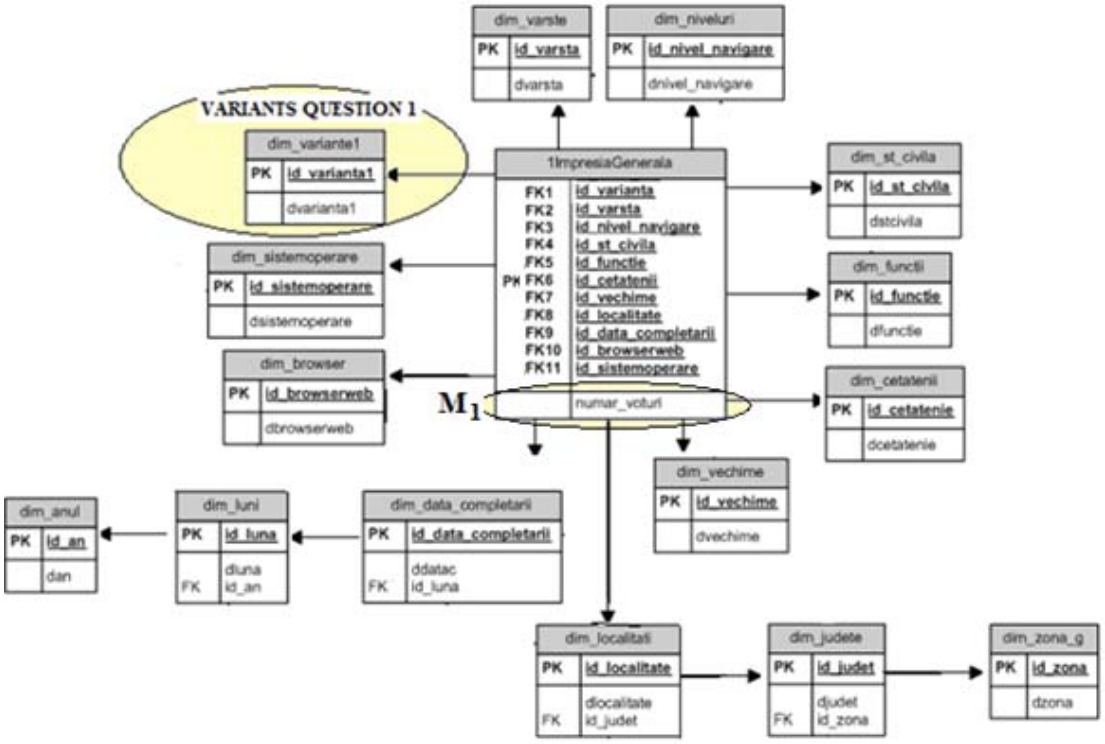


Fig. 10. Sub-schema of the data mart. Analyzing the first question

DNIVEL_NAVIG...	DVARSTA	DSTCIVILA	DFUNCTIE	DCET...	DVECHIME	DLOCALITATE	DDATAC	DBROWSER	DSISTE...	DVARIANTA1	NUMAR_VOT...
bun cunoscator	intre 19 si 23...	necasat...	elev/student	romana	0 luni	Timisoara	22-10-2012	Google Chrome	Windows	bun	2
bun cunoscator	intre 19 si 23...	necasat...	elev/student	romana	0 luni	Timisoara	30-10-2012	Google Chrome	Windows	foarte bun	1
bun cunoscator	intre 19 si 23...	necasat...	elev/student	romana	sub 1 an	Dr.Tr.Severin	18-10-2012	Mozilla Firefox	Windows	bun	1
bun cunoscator	intre 35 si 49...	necasat...	profesor	romana	peste 10 ani	Timisoara	18-10-2012	Google Chrome	Windows	bun	1
expert	intre 19 si 23...	necasat...	elev/student	romana	0 luni	Timisoara	19-10-2012	Mozilla Firefox	Windows	foarte bun	1
expert	intre 24 si 34...	necasat...	manager	romana	intre 5 si 10...	Timisoara	20-10-2012	Google Chrome	Windows	destul de bun	1
bun cunoscator	intre 19 si 23...	necasat...	elev/student	romana	sub 1 an	Timisoara	20-10-2012	Mozilla Firefox	Windows	bun	1
expert	intre 19 si 23...	necasat...	elev/student	romana	sub 1 an	Sofronea	20-10-2012	Google Chrome	Windows	bun	1
expert	intre 19 si 23...	necasat...	specialist IT	romana	sub 1 an	Timisoara	21-10-2012	Mozilla Firefox	Windows	destul de bun	1
bun cunoscator	intre 19 si 23...	necasat...	elev/student	romana	0 luni	Timisoara	22-10-2012	Google Chrome	Windows	foarte bun	2
expert	intre 19 si 23...	divortat	economist	alta ...	intre 1 si 3 ani	Timisoara	22-10-2012	Google Chrome	Windows	bun	1
expert	intre 19 si 23...	necasat...	economist	romana	sub 1 an	Timisoara	23-10-2012	Google Chrome	Windows	bun	1
incepator	sub 15 ani	necasat...	elev/student	romana	intre 5 si 10...	Oradea	23-10-2012	Internet Explorer	Windows	foarte bun	1
expert	intre 24 si 34...	necasat...	altceva	romana	intre 3 si 5 ani	Timisoara	28-10-2012	Google Chrome	Windows	destul de bun	1
bun cunoscator	intre 24 si 34...	necasat...	specialist IT	romana	sub 1 an	Timisoara	28-10-2012	Opera	Windows	destul de bun	1
bun cunoscator	intre 19 si 23...	necasat...	elev/student	romana	intre 1 si 3 ani	Timisoara	17-10-2012	Mozilla Firefox	Windows	destul de bun	1
bun cunoscator	intre 19 si 23...	necasat...	elev/student	romana	0 luni	Ineu	18-10-2012	Google Chrome	Windows	bun	1
bun cunoscator	intre 19 si 23...	necasat...	elev/student	romana	0 luni	Satu Mare	18-10-2012	Google Chrome	Windows	bun	1
bun cunoscator	intre 19 si 23...	necasat...	elev/student	romana	sub 1 an	Timisoara	19-10-2012	Google Chrome	Windows	foarte bun	1
bun cunoscator	intre 19 si 23...	necasat...	elev/student	romana	sub 1 an	Timisoara	20-10-2012	Opera	Windows	foarte bun	1
incepator	intre 24 si 34...	necasat...	economist	romana	intre 5 si 10...	Valcea	20-10-2012	Mozilla Firefox	Windows	bun	1
bun cunoscator	intre 19 si 23...	necasat...	elev/student	romana	0 luni	Marga	20-10-2012	Google Chrome	Windows	destul de bun	1

Fig. 11. Visualization of the integrated data

The predominant response for the first question is that the general impression regarding our faculty's web site is good or very good (79%).
In the following analysis, the dimensions took into account are: the preferred operating system and the preferred browser, with the alternatives: very good, good, pretty good and poor. The responses can be easily depicted as a list or as a chart – Figure 12.
Most of the respondents use Windows as an operating system (97%), 2 respondents use IOS and 2 respondents prefer Android. From the Windows users, 101 respondents out of 131 think that our faculty's web site is very good or pretty good, more than 78%.
From the majority „Windows users", 68 respondents prefer the web browser Google Chrome, this being the overall preferred choice.

Script Output | Query... | SQL | All Rows Fetched: 17 in 0,075 seconds

	DSISTEMOPERARE	DBROWSER	DVARIANTA1	NUMAR_VOTUR
1	iOS	Google Chrome	bun	
2	Windows	Internet Explorer	foarte bun	
3	iOS	Mozilla Firefox	bun	
4	Windows	Google Chrome	destul de bun	1
5	Windows	Internet Explorer	destul de bun	
6	Windows	Opera	destul de bun	
7	Windows	Mozilla Firefox	foarte bun	1
8	Windows	Mozilla Firefox	bun	2
9	Windows	Opera	foarte bun	
10	Windows	Mozilla Firefox	destul de bun	
11	Windows	Opera	bun	
12	Windows	Mozilla Firefox	slab	
13	Android	Google Chrome	destul de bun	
14	Android	Mozilla Firefox	slab	
15	Windows	Internet Explorer	bun	
16	Windows	Google Chrome	bun	3
17	Windows	Google Chrome	foarte bun	1

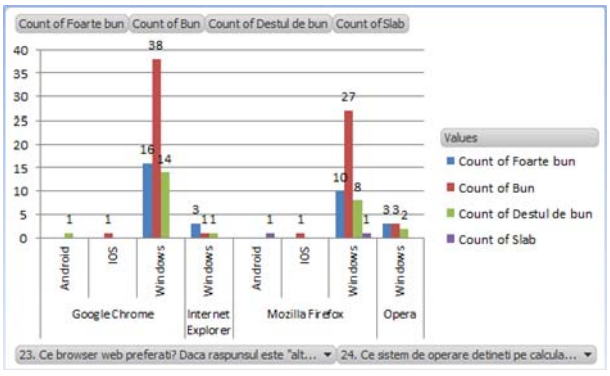


Fig. 12. Responses for question1, the general impression of the web site, taking into consideration the preferred operating system and the preferred browser

Regarding the displaying of commercial advertisements, 62% of the respondents think that this is a good choice. The majority of the respondents are still students (70%) and the result regarding the necessity of commercial advertisement is in agreement with our

policy – Figure 13. Because it is considered a quick way to visualize the analysis, forceful (it emphasized the main point), convincing (proves a point) and because of the enhanced flexibility, results were displayed as charts from now on.

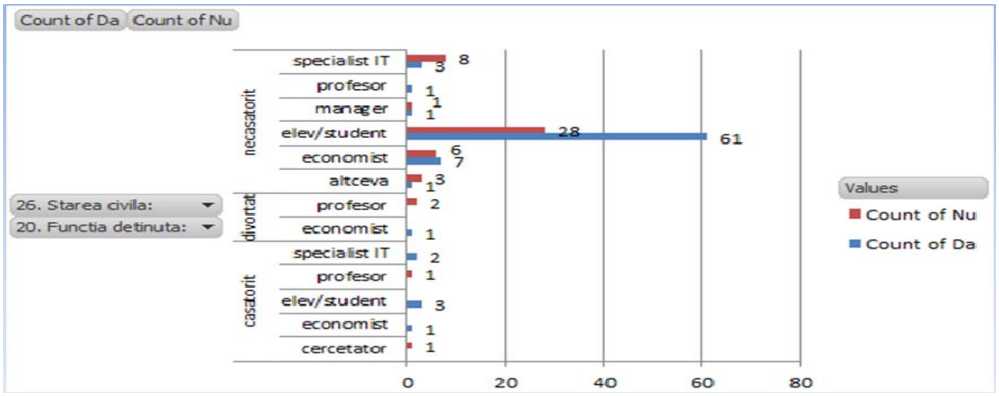


Fig. 13. Responses for question 6, the necessity of displaying commercial advertisements, taking into consideration the marital status and the position

The following analysis regards **whether** or not parents should have access to the student's records, and the result is in favor (60%). The dimensions took into account are the age and the residence geographical area. Most of the respondents live in Banat (101 persons), 41 of them with ages between 19 and 23 – Figure 14. From the Banat respondents, 58% think that the parents should have access to their academic records. The geographical location hierarchy was used, rolling up from the city in order to display the geographical area. There were no respondents with the ages between 15 and 18 in the time slot took into consideration.

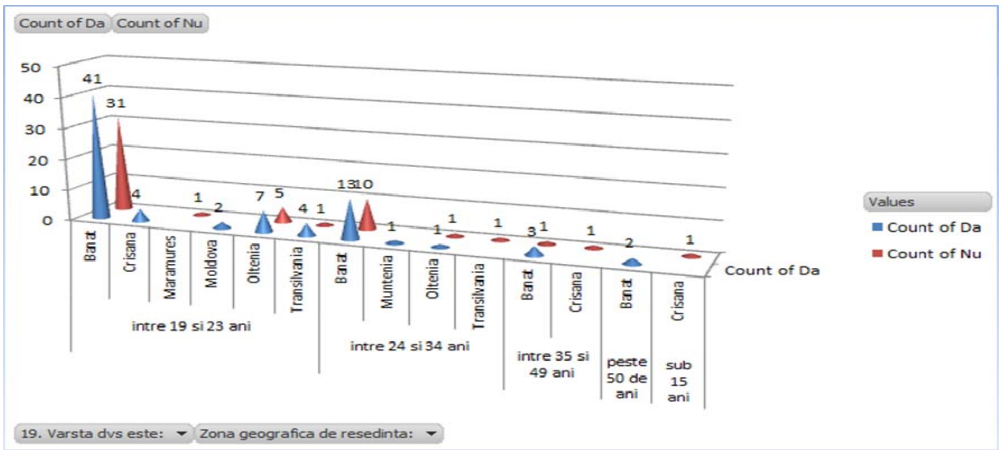


Fig. 14. Responses for question 17, weather or not parents should have access to the students records, taking into consideration the age and the geographical area

The questionnaire was completed during two school weeks. The first day was 14th of October and the last day 30th of October. Some days like 19th of October 2012 were more productive than the others (34% of the total responses). The following analysis is about the time a respondent spend on our faculty's web site. The result shows that most of the respondents usually spend between 0 and 15 minutes on our web site (67%), or between 15 and 30 minutes (28%) – Figure 15.

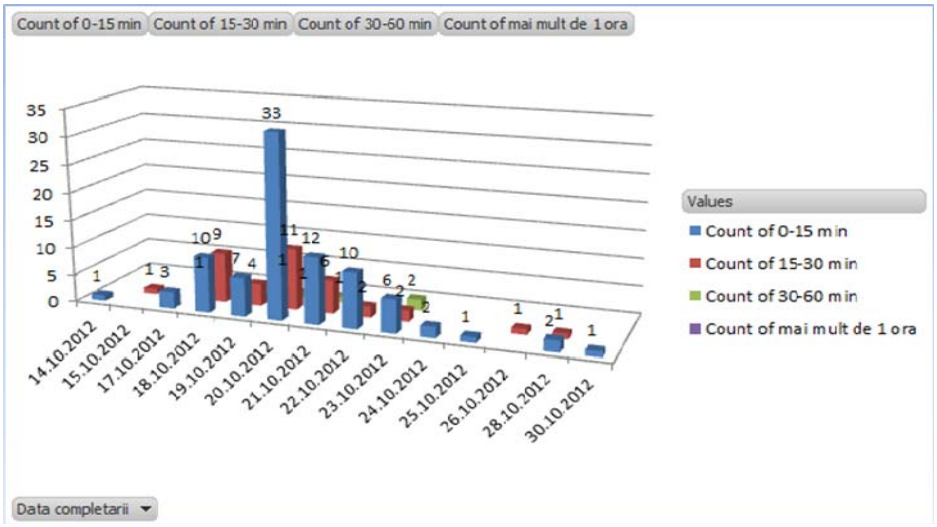


Fig. 15. Responses for question 9, the time that the respondent spent browsing for information on the web site, taking into consideration the date when he/she completed the questionnaire

55% of our respondents want to be informed using the University's Press Magazine. From the Romanian citizenship with good browsing the web skills, 52 want to be subscribed to our University's Press Magazine and 36 don't. At the expert level, the result is irrelevant – figure 16.

Regardless of the citizenship, most of the respondents have good navigation skills (70%) and 28% of them consider themselves experts.

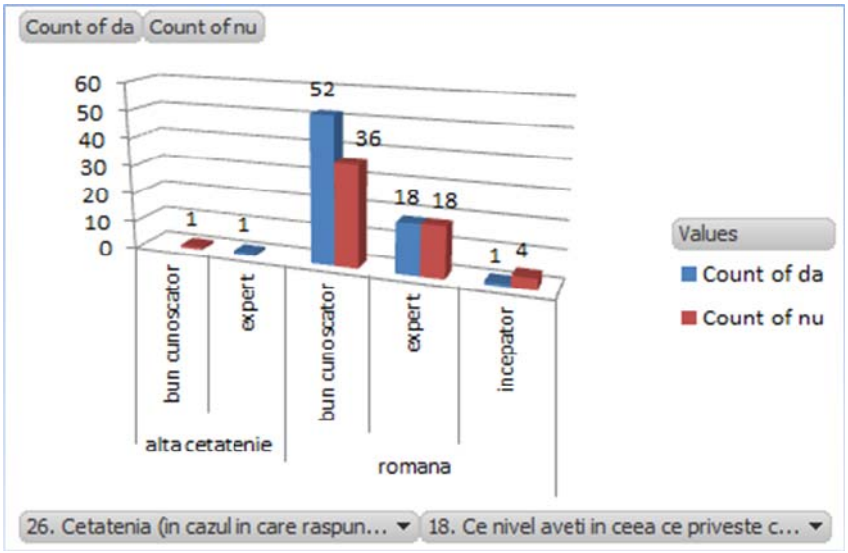


Fig. 16. Responses for question 13, if they want to be subscribed to the University's Press Magazine, taking into consideration the citizenship and the browsing the web skills

5 Conclusions

Questionnaires are common used data gathering tools for collection of primary data in surveys, analysis or to corroborate other research findings. Analyzing the data is a task that requires good analytical skills, the results and conclusions implying further decisions and strategies.

Base data analysis can be made i.e. with MS Excel Data Analysis – Descriptive Statistics and Pivot Table, but to obtain more value from the data, statistical software i.e. SPPS is recommended. Detailed, valuable information about the attitudes and behavior of the questionnaire respondents is extracted.

Beyond these traditional approaches, the data collected through a questionnaire can be transposed into a multidimensional data model. The model is built around measures that are aggregated according to the introduced dimensions. Time and/or spatial dimensions are typically dimensions within the model. According to the identified multidimensional model, a corresponding data mart will enable further advanced analyses. The proposed agile development framework enables the rapid deployment of the data mart taken into consideration its environment.

Additionally, a study case was proposed, a questionnaire built and different analyses presented.

Further research might go deeper into modeling the data mart environment, enabling various views of the collected data and extending the data analysis capabilities.

References

[1] Lastrucci C. L., *The Scientific Approach: Basic Principles of the Scientific Method*, Schenkman Pub. Co., 1967

[2] Foltean F., *Cercetări de marketing*, Editura Mirton, Timi□oara, 2000

[3] V. Jupp, *The SAGE Dictionary of Social Research Methods*, SAGE Publishing Ltd., 2006

[4] M. Muntean, *Sisteme pentru asistarea deciziilor*, Suport de curs on-line, 2010

[5] A. Williams, *How to… Write and Analyze a Questionnaire*, British Orthodontic Society White Paper, 2003

[6] I. Plăia□, *Cercetări de marketing*, Editura Risoprint, Cluj-Napoca, 2008

[7] McKnights, W., The New Business Intelligence Architecture Discussion, *Information Management Magazine*, September 2004

[8] Lungu, I.(coord), Bara A, Bodea C, Botha I., *Tratat de baze de date. Organizare, proiectare, implementare*, Editura ASE, 2011

[9] A.N. AbuAli and H.Y. Abu-Addose, "Data Warehouse Critical Success Factors," *European Journal of Scientific Research*, vol. 42, no. 2, pp. 326 - 335, 2012

[10] C. Adamson, *Mastering Data Warehouse Aggregates: Solutions for Star Schema Performance*. Indianapolis, Indiana: Wiley Publishing, Inc., 2006

[11] C. Ballard, D.M. Farrell, A. Gupta, C. Mazuela, and S. Vohnik, *Dimensional Modeling: In a Business Intelligence Environment*, 1st ed.: IBM Corporation, Redbooks, 2006

[12] J.B. Barlow et al., "Overview and Guidance on Agile Development in Large Organizations," *Communications of the Association for Information Systems*, vol. 29, 2011

[13] M. Golfarelli, D. Maio, and S. Rizzi, "The Dimensional Fact Model: a Conceptual Model for Data Warehouses," *International Journal of Cooperative Information*, vol. 7, no. 2, pp. 215 - 247, 1998

[14] M. Nagy, "A Framework for Semi-Automated Implementation of Multidimensional Data Models,"

Database Systems Journal, vol. 3, no. 2, pp. 31-40, July 2012

[15] N. Rahman, D. Rutz, and S. Akher, "Agile Development in Data Warehousing," *International Journal of Business Intelligence Research*, vol. 2, no. 3, pp. 64-77, July-September 2011, DOI: 10.4018/jbir.2011070105

[16] T. Spencer and T. Loukas. (1999, Jan.) *From Star to Snowflake to ERD: Comparing Data Warehouse Design Approaches*

[17] E. Malinowski, E. Zimányi, Hierarchies in a multidimensional model: From conceptual modeling to logical representation, *Data & Knowledge Engineering,* 59 (2006) 348–377 http://code.ulb.ac.be/dbfiles/MalZim2006article.pdf

[18] B. H. Wixom, H. J. Watson, An empirical investigation of the factors affecting data warehousing success, *Journal MIS Quaterly*, Volume 25 Issue 1, March 2001, pages 17-32

[19] T.B. Pedersen, C.S. Jensen, C.E. Dyreson, A foundation for capturing and querying complex multidimensional data, *Information Systems*, volume 26, issue 5, July 2001, pages 383–423

[20] J.M. Perez, R. Berlanga, M.J. Aramburu; T.B. Pedersen, Integrating Data Warehouses with Web Data: A Survey, *IEEE Transactions on Knowledge and Data Engineering*, July 2008, volume 20, page(s): 940 – 955

[21] T.B. Pedersen, A foundation for capturing and querying complex multidimensional data, *Information Systems*, Volume 26, number 5, July 2001 , pp. 383-423

[22] M. Nagy, *Design and Implementation of Data Warehouses for Business Intelligence applied in Business*, Doctoral Thesis, Cluj-Napoca, 2012

[23] M. Muntean, D. Târnăveanu, A. Paul, BI Approach for Business Performance, *Proceedings of the 5th WSEAS Conference on Economy and Management Transformation*, 2010

With a background in Computer Science and a Ph.D. obtained both in Technical Science and in Economic Science (Economic Informatics), professor **Mihaela I. MUNTEAN** focused her research activity on topics like knowledge management, business intelligence, business information systems, information system management. Over 70 papers in indexed reviews and conference proceedings and the involvement with success in 7 multi-annual national research grants/projects are sustaining her contributions in the research fields mentioned above. Currently, professor Mihaela I. Muntean is the chair of the Business Information Systems Department at the West University of Timişoara and an IT independent consultant.

Diana TÂRNĂVEANU has graduated the Faculty of Mathematics from the West University of Timişoara in 1995. She holds a PhD diploma in Management from 2008. Currently she is a lecturer of Economic Informatics within the Department of Business Information Systems at Faculty of Economics and Business Administration from the West University of Timişoara. She is teaching database management systems, business intelligence and decision support systems. She is the author of more than 10 books and over 35 journal articles in the field of Knowledge Management, Decision Support Systems, Collaborative Systems and Business Intelligence.

Multi-level and Multi-component Bitmap Encoding for Efficient Search Operations

Madhu BHAN[1] , Dr.RAJANIKANTH K [2], Dr. Suresh KUMAR T.V[3]
[1,2]M.S.Ramaiah Institute of Technology, Department of Computer Applications, India.
[3]Visvesvaraya Technological University, India
madhoobhan@yahoo.co.in

The growing interest in data warehousing for decision makers is becoming more and more crucial to make faster and efficient decisions. On-line decision needs short response times. Many indexing techniques have been created to achieve this goal in read only environments. Indexing technique that has attracted attention in multidimensional databases is Bitmap Indexing. The paper discusses the various existing bitmap indexing techniques along with their performance characteristics. The paper proposes two new bitmap indexing techniques in the class of multi-level and multi-component encoding schemes and prove that the two techniques have better space–time performance than some of the existing techniques used for range queries. We provide an analytical model for comparing the performance of our proposed encoding schemes with that of the existing ones.

Keywords: *Bitmap encoding, Datawarehouse, multi-level indexing, multi-component indexing, On-Line Analytical Processing.*

1 Introduction

While the query performance issues of on-line transaction processing (OLTP) systems have been extensively studied and are pretty much well-understood, the state-of-the-art for data warehouse systems is still evolving as indicated by the growing active research in this area [1]. In particular, Data warehouse systems operate in read-mostly environments, which are dominated by complex adhoc queries that have high selectivity factors [2]. Due to large size of the data warehouse and the complexity of queries, quick response time plays an important role as timely access to information is the basic challenge to match the pace of the query results with the speed of thought of the user. From various methods available to improve performance, indexing ranks very high [3]. Indexes are database objects associated with database tables and created to speed up access to data within table. Index space and access time play an important role in choosing an indexing technique in data warehouse. If the space used by an index is large then the results are achieved in short time on the other hand if the space used by the index space is small then the results are achieved in greater amount of time. So there is a trade-off between the time consumed and the space used by a particular index. A committed approach to answer complex queries swiftly in Data warehouse systems is the use of bitmap indexing [4], [5] and [6]. Bitmap manipulation techniques have already been used in some commercial products [7] to speed up query processing. The basic bitmap index uses each distinct value of the indexed attribute as a key, and generates one bitmap containing as many bits as the number of records in the data set for each key [8]. The advantage of bitmap index is that complex selection predicates can be computed very quickly by performing bit-wise AND, OR and NOT operations on bitmap indices. Bitmaps are

well supported by hardware and are easy to compress. Each individual bitmap is small and frequently used ones can be cached in memory. This property of bitmap has led to considerable interest in their use in Decision Support systems. The size of a basic bitmap index is relatively small for low-cardinality attributes, such as "gender," "types of cars sold per month," or "airplane models produced by Airbus and Boeing." However, for high-cardinality attributes such as "temperature values in a supernova explosion," the index sizes may be too large to be of any practical use [9]. In the literature, there are three basic strategies to reduce the sizes of bitmap indices: (1) using more complex bitmap ***encoding*** methods to reduce the number of bitmaps or improve query efficiency, (2) ***compressing*** each individual bitmap, and (3) using ***binning*** or other mapping strategies to reduce the number of keys. Various bitmap indexes have been designed for different query types, including range queries, aggregation queries, and OLAP-style queries. However, as there is no overall best bitmap index over all kinds of queries, maintaining multiple types of bitmap indexes for an attribute may be necessary in order to achieve the desired level of performance. While the gains in query performance using a multiple-index approach might be offset by the high update cost in OLTP applications, this is not an issue in the read-mostly environment of data warehouse applications. In the remaining of this paper, we first present in Section 2 a review of different bitmap indexing strategies. We discuss the three basic encoding techniques namely Equality encoding, Range encoding and Interval encoding along with their performance characteristics in Section 3. In Section 4 and 5, the proposed multi-level encoding and multi-component encoding technique is defined with an analytical model. Section 5 concludes the paper with future enhancements of the proposed techniques.

2 Related work.

Various bitmap indexes have been demonstrated to significantly speed up searching operations in data warehousing, On-Line Analytical Processing (OLAP), and many scientific data management tasks [10] and [11 . This has led a number of commercial database management systems (DBMS) to support bitmap indexes [7]. However, most of the bitmap index implementations in commercial DBMS are relatively simple, such as the basic bitmap index or the bit-sliced index. There is a significant number of promising techniques proposed in the research literature that have not gained wide acceptance yet. A bitmap index typically uses a combination of three types of strategies namely encoding, binning and compression, though it is common to omit one or two. For example, the first commercial implementation of a bitmap in Model 204 uses equality encoding without binning or compression [8].

- **Encoding**: In the simplest encoding, a bitmap corresponds to exactly one attribute value. This encoding is known as Equality encoding where i-th bit is set to 1 if the i-th row of the base table has a value for the indexed column. It is possible to reduce the number of bitmaps by using a different encoding method. Fig 1(a) shows an example of encoded bitmap index. Assume that the attribute domain given by the table T is{a, b, c}. A simple bitmap index uses four bitmap vectors whereas an encoded bitmap index uses $\lceil \log_2 3 \rceil = 2$ bitmap vectors plus a mapping table. It encodes the values from a simple bitmap index by means of Huffman encoding. Thus we see that for an attribute with C distinct values we use only $\log_2 C$ encoded bitmap vectors instead of C bitmap vectors. We assume that we have a fact table SALES with N tuples and a dimension table PRODUCT with 12,000

different products. If we build a simple bitmap index on PRODUCT, It will require 12,000 bitmap vectors of N bits in length. However, if we use encoded bitmap indexing we only need $\lceil \log_2 12{,}000 \rceil = 14$ bitmap vectors plus a mapping table. It is a very significant reduction of the space complexity. Other common encoding schemes include range encoding and interval encoding [2]. More sophisticated encoding schemes can be generated from the above three basic encoding schemes, Equality encoding, Range encoding and Interval encoding. One approach of extending the basic encoding schemes is the multi-level encoding which can be viewed as using hierarchy of levels with different encoding techniques. Another strategy is to decompose the attribute value into several components and encode each component using a basic encoding scheme. These are called as multi-component encodings [12] and [13]. The best known example of such an encoding is the binary encoding which is also known as the bit-sliced index. Finding the optimal encoding method that balances query performance and index size remains an interesting challenge.

- **Compression:**

Compressing each bitmap in a bitmap index can save space. A lossless compression method can be used for this purpose. There has been considerable amount of work done on this subject [14] and [15] . For example, most generic text compression methods, such as LZ77, are effective in reducing the index size on disk, but they can also significantly increase the time required to answer a query, because the compressed bitmaps have to decompressed before being used in logical operations. Bitmap compression algorithms typically employ run-length encoding such as the Byte-aligned Bitmap Code and the Word-Aligned Hybrid code [16]. These compression methods require very little effort to compress and decompress. The Byte-Aligned Bitmap Code (BBC) can compress bitmaps and at the same time it also reduces the query response time. The BBC compressed basic bitmap index is implemented in ORACLE DBMS. The Word-Aligned Hybrid (WAH) code has been shown to outperform BBC in most cases[18]. This method trades some space for more efficient CPU operations. In one set of tests, it was shown to use about 50% more space than BBC, but answered queries 10 times faster on average More importantly, bitmaps compressed with WAH, BBC, PLWAH and CONCISE can directly participate in bitwise operations without decompression. This gives them considerable advantages over generic compression techniques such as LZ77. Fig1 (b) shows an WAH bit vector representing 128 bits. Assuming that computer word length is 32 bits, each literal word stores 31 bitmaps from the bitmap and each fill word represents a fill with a multiple of 31 bits. The second line in Figure shows how the bitmap is divided into 31-bit groups and the third line shows the hexadecimal representation of the groups. The last line shows the values of the WAH words. The logical operations can be directly performed on the compressed bitmaps and the time needed by one such operation on two operands is related to the sizes of the compressed bitmaps. Extended work on WAH compression has shown further improvements in performance of query processing. Different compression methods can be used and each may have a drastically different query processing costs.

- **Binning:**

In Binning, bitmap indices are built on attribute ranges rather than on distinct attribute values. For high-cardinality columns, it is useful to bin the values, where each bin covers multiple values and build the bitmaps to represent the values in each

bin [17] .The advantage of this approach is that a lower number of bitmap vectors is required. On the other hand, parts of the original data (candidates) have to be read from disk in order to answer the queries correctly. This process is called candidate check . This approach reduces the number of bitmaps used regardless of encoding method. However, binned indexes can only answer some queries without examining the base data. An example of a bitmap index with bins is given in Figure 1(c). Assume that we want to evaluate the query 37 <= x < 63. Bins 1, 2 and 3 contain the relevant data values. The bin in which a query boundary falls is known as an edge bin. Thus bins 1 and 3 are edge bins since they contain also irrelevant values, answering this query involves checking the values on disk corresponding to the four "1-bits" in these two columns. In this example only one of the four values qualifies, namely, 61. We call this additional step the candidate check. As we can see from this example, the cost of performing a candidate check on an edge bin is related to the number of "1-bits" in that bin. The process of checking the base data is known as the candidate check. In most cases, the time used by the candidate check is significantly longer than the time needed to work with the bitmap index [18] and [19]. Therefore, binned indexes exhibit irregular performance. They can be very fast for some queries, but much slower if the query does not exactly match a bin. The key advantage of binning is that it may reduce index sizes

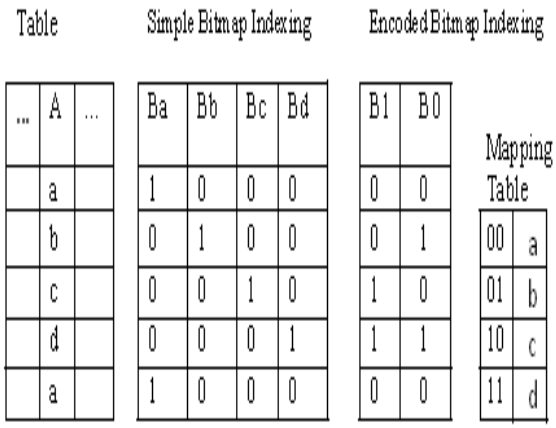
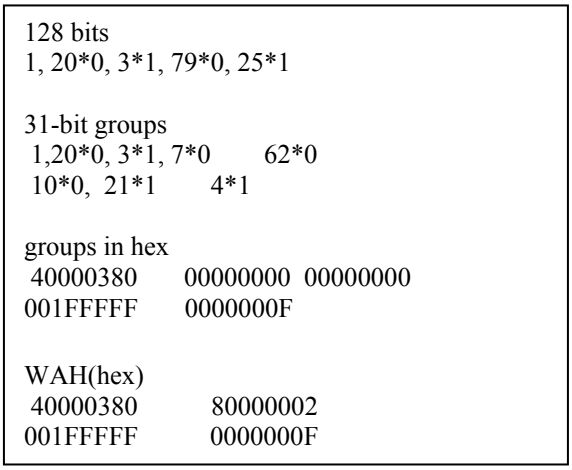


Fig 1(a)



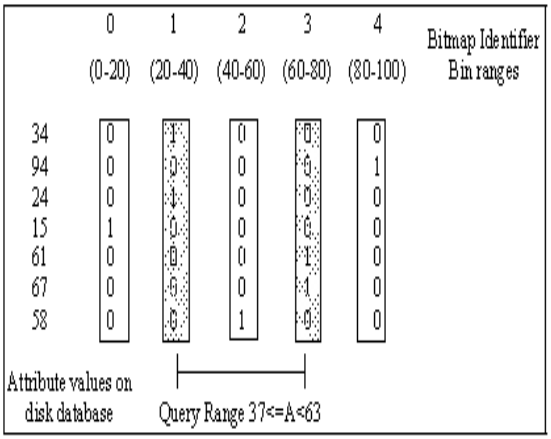


Fig. 1(c)

however, the disadvantage is that the index is no longer able to fully resolve all queries Among the three, encoding is by far the largest category, however their impacts on the overall index performance are less

studied than those of binning or compression. For this reason, studying the encoding methods is more likely to lead to the best bitmap index method.

3 Comparison of basic Encoding Schemes

We first review the three existing bitmap encoding schemes, equality encoding, denoted by E, range encoding denoted by R and interval encoding denoted by I. These schemes have been described in several papers under different names [1], 2] and [13]. Equality encoding is the most fundamental and common bitmap encoding scheme. It consists of C bitmaps E = {EO, El,. . . ,EC-1}, where each bitmap Ev = {v}. Consider an attribute A of a relation R, where the attribute cardinality is c. For simplicity and without loss of generality, the domain of A is assumed to be a set of consecutive integers from 0 to C - 1. This allows us to use set operators and logical operators interchangeably. The logical operators AND, OR, and XOR are denoted by Λ, V, and ⊖ respectively. For example, Figure 2(a) shows the Equality-encoded bitmap index. The leftmost column shows the row ids (RID) for the data values represented by the projection index on an attribute A with cardinality C = 10 of a 12 record relation R. Figure 2(b) shows the Equality encoded bitmap index for the data in Figure 2(a), where each column represents an equality-encoded bitmap Ev associated with an attribute value v. This strategy is the most efficient for equality queries such as "A = 3 which needs only one bitmap E2 to be accessed. The Range encoding and Interval encoding techniques are optimized for one-sided and two-sided range queries, respectively. An example of a one-sided range query (1RQ) is "A <= 3". A two-sided range query (2RQ), for instance, is "6 < A < 8". A comparison of an Equality encoding, Range encoding and Interval encoding is given in Figure 2 . Let us look at the encoding of value 2, in Equality encoding and Range encoding first, which is highlighted in the Figure2(a) and Figure 2(b). For Equality encoding, the third bitmap is set to "1" (E2), whereas all other bits on the same horizontal line are set to "0". For the Range-encoded bitmap index, all bits between bitmap R2 and R8 are set to "1", the remaining bits are set to "0". Range encoding is very efficient for evaluating range queries. Consider, for instance, the query "A <= 4". In this case, at most one bitmap, namely bitmap R4, has to be accessed (scanned) for processing the query. All bits that are set to "1" in this bitmap fulfil the query constraint. On the other hand, for the Equality encoded bitmap index, the bitmaps E0 to E4 have to be ORed together (via the Boolean operator OR). This means that, Range encoding requires at most one bitmap scan for evaluating range queries, whereas Equality encoding requires in the worst case C/2 bitmap scans, where C corresponds to the number of bitmaps. Since one bitmap in Range encoding contains only "1"s, this bitmap is usually not stored. Therefore, there are C-1 bitmaps in a range-encoded index. The Interval encoding, I, is optimal for the two sided Range queries. In Range encoding, each bitmap Ri = [0, i], and each 2RQ -query is evaluated by

	$\pi_A(R)$	E^9	E^8	E^7	E^6	E^5	E^4	E^3	E^2	E^1	E^0
1	3	0	0	0	0	0	0	1	0	0	0
2	2	0	0	0	0	0	0	0	1	0	0
3	1	0	0	0	0	0	0	0	0	1	0
4	2	0	0	0	0	0	0	0	1	0	0
5	8	0	1	0	0	0	0	0	0	0	0
6	(2)	0	0	0	0	0	0	0	(1)	0	0
7	9	1	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	1
9	7	0	0	1	0	0	0	0	0	0	0
10	5	0	0	0	0	1	0	0	0	0	0
11	6	0	0	0	1	0	0	0	0	0	0
12	4	0	0	0	0	0	1	0	0	0	0

Fig 2 (a)

	$\pi_A(R)$	R^8	R^7	R^6	R^5	R^4	R^3	R^2	R^1	R^0
1	3	1	1	1	1	1	1	0	0	0
2	2	1	1	1	1	1	1	1	0	0
3	1	1	1	1	1	1	1	1	1	0
4	2	1	1	1	1	1	1	1	0	0
5	8	1	0	0	0	0	0	0	0	0
6	(2)	1	1	1	1	1	1	0	0	0
7	9	0	0	0	0	0	0	0	0	0
8	0	1	1	1	1	1	1	1	1	1
9	7	1	1	0	0	0	0	0	0	0
10	5	1	1	1	1	0	0	0	0	0
11	6	1	1	1	0	0	0	0	0	0
12	4	1	1	1	1	1	0	0	0	0

Fig 2(b)

	$\pi_A(R)$	I^4	I^3	I^2	I^1	I^0
1	3	0	1	1	1	1
2	2	0	0	1	1	1
3	1	0	0	0	1	1
4	2	0	0	1	1	1
5	8	1	0	0	0	0
6	2	0	0	1	1	1
7	9	0	0	0	0	0
8	0	0	0	0	0	1
9	7	1	1	0	0	0
10	5	1	1	1	1	0
11	6	1	1	1	0	0
12	4	1	1	1	1	1

Fig 2(c)

operating on an appropriate pair of bitmaps: [x, y] = R^y (+) R^{x-l}. The Interval encoding scheme based on range encoding consists of [C/2] bitmaps I = $\{I^0, I^1, I.^2 \ . \ . \ , I^{[C/2-1]}$, where each bitmap I^j =[j, j + m], and m = [C/2] - 1. The Interval encoding, I, is optimal for the two sided range queries. Figure 2(c) shows the Interval encoded bitmap index for the data in Figure 2(a). A 2RQ, in general, is evaluated by operating on a pair of bitmaps: [x,y] = $I^x \Lambda I^{y-m}$. The Interval-encoding scheme[2] reduces the number of bitmaps only by a factor 2 while still guaranteeing at most a two-scan evaluation for any query . Thus, other techniques are needed to make bitmap indices practical for high cardinality attributes [9]. The encoding method that produces the least number of bitmaps is Binary encoding. This encoding method uses only $\log_2 C$ rather than C/2 bitmaps, where C is the attribute cardinality. The advantage of this encoding is that it requires much fewer bitmaps than Interval encoding. However, to answer a range query, using interval encoding one has to access only two bitmaps whereas using binary encoding one usually has to access all bitmaps. An Equality encoded index may access a large number of bitmaps to answer a Range query, but the bitmaps are usually relatively easy to compress, while the Range encoding and the Interval encoding access fewer bitmaps to answer a range query, but they produce bitmaps that are hard to compress. Equality encoding requires C bitmaps, Range encoding requires C-1 bitmaps and Interval encoding reduces the number of bitmaps only by a factor of 2. Controlling the size of bitmap indicies is crucial to make bitmap indicies practical for high cardinality attributes. We find that Equality encoded index accesses larger number of bitmaps to answer a query but here it is easy to compress the bitmaps, while Range encoding and Interval encoding access few bitmaps ,but the bitmaps produced are hard to compress. Therefore, it is worthwhile to explore strategies that combine the advantages of the three basic bitmap indexing techniques. Strategies like multi-level encoding and multi-component encoding have been proposed by authors to reduce the index size and bitmap scans.

4 Multi-level Encoding

We find that Equality encoded index accesses larger number of bitmaps to answer a query but it is easy to compress the bitmaps, while Range encoding and Interval encoding access few bitmap , but the bitmaps produced are hard to compress

[15],[16]. To combine the advantages of these encoding techniques we can explore multilevel encoding techniques. We can think of a multi-level encoding as encoding at multiple levels where each level can be encoded separately using any encoding method. In previous works [6] we have seen that the multi-level encoding, when used with binning methods require candidate checks, which resulted in performance measurements that do not truly represent the characteristics of the encoding schemes. In this paper, we study the multi-level encoding with binning, which removes the need for candidate checks. This allows a better understanding of the performance characteristics of these multi-level encodings. We take an analytical approach in our study which allows us to compare various parameters like number of bins and cardinality of attributes.

- **Methodology**

Conceptually we can think of our multilevel bitmap technique as formed of equality encoding with binning in the first level and followed by binary encoding at the second level. Figure 3 shows an example of our multi-level encoding techniques with the same attributes values as given in the above example(Figure 1). To answer a query we first scan the equality encoded bitmaps(E_{0-1},E_{2-3},E_{3-4}……,E_{8-9}). Based on these bitmaps we need to scan Binary bitmap vector B. Consider for instance the Query A = 3. Our multi-level coding first access bitmap E_{2-3}. Then corresponding to 1's in E_{2-3} it scans binary bitmap B. The 1's in B indicate 3 and 0's in B indicate 2 as shown in mapping table. Thus we see that with C =10 and bin size n =2 we require 5(C/2) + 1 bitmap vectors scans.

1. Equality Encoding

Case of equality queries(A=3) :-

TIME: One bitmap E^2 needs to be scanned.

SPACE: 10(C) bitmaps are formed.

Case of 1RQ(A<=3) :-

TIME: Four bitmaps E^0 to E^3 need to be scanned and ORed.

SPACE:10(C) bitmaps are formed.

2. Our Multilevel Encoding:

Case of Equality queries(A=3):-

TIME: One bitmap E_{2-3} + Binary bitmap B needs to be scanned.

SPACE: 5(C/n) bitmaps where n is the bin size + 1Binary Bitmap is formed + mapping table

Case of One sided Range Query(A<=3):-

TIME: Two bitmaps $E_{0-1,}$ E_{2-3} need to be scanned and ORed.

SPACE:5(C/n) bitmaps where n is the bin size +1 Binary bitmap is formed + mapping table.

From the above we analyse that our proposed multi-level encoding technique is suitable for 1-sided range queries where the index space as well as number of bitmaps scanned is reduced by a factor of n. The above analyses is true for one sided range queries where query condition contains the upper bound value of the specific bin. However if the query condition contains the lower value of the specific bin additional $\log_2 n$ binary vectors need to be scanned. In such cases also the number of scans for one sided range queries is less than that required in Equality Encoding. For the given example the comparison between no. of scans for different query conditions of range queries is given in Figure 4. For range query A <= 6, Equality encoding requires eight bitmaps to be scanned whereas new multi-level encoding requires only four equality encoded bitmaps $E_{0-1,}$ $E_{2-3,}$ $E_{4-5,}$ E_{6-7} and Binary bitmap B to check for corresponding 0's which indicate value 6 so that value 7 is eliminated.

RID	$\Pi_A(R)$	E_{8-9}	E_{6-7}	E_{4-5}	E_{2-3}	E_{0-1}	B
1	3	0	0	0	1	0	1
2	2	0	0	0	1	0	0
3	1	0	0	0	0	1	1
4	2	0	0	0	1	0	0
5	8	1	0	0	0	0	0
6	2	0	0	0	1	0	0
7	9	1	0	0	0	0	1
8	0	0	0	0	0	1	0
9	7	0	1	0	0	0	1
10	5	0	0	1	0	0	1
11	6	0	1	0	0	0	0
12	4	0	0	1	0	0	0

Mapping Table

0 2 4 6 8	0
1 3 5 7 9	1

Fig 3. Multi-level encoding techniques

- **Analytical Model**

This section will compare the space-time trade off of the aforementioned bitmap encoding schemes**.** Let C denote the Cardinality and n denote the bin size.

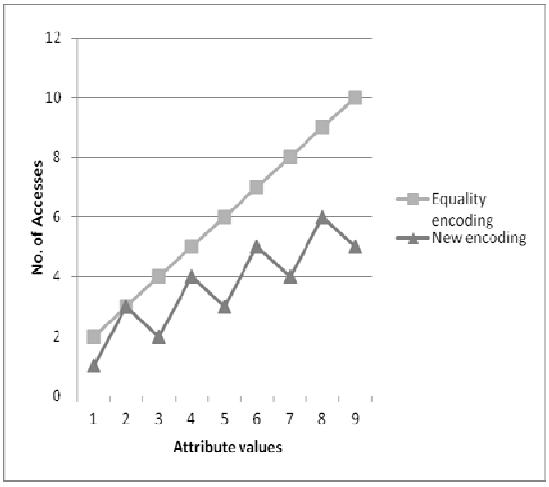


Fig 4. The comparison between no. of scans for different query conditions of range queries

Definition of cardinality in set theory refers to the number of members in the set. On database theory, the cardinality of a table refers to the number of rows contained in a particular table. In terms of OLAP system, cardinality refers to the number of rows in a table. On the other hand, on a data warehousing point of view, cardinality usually refers to the number of distinct values in a column. We compare the values of space and time at different values of C(10,100,1000). For an average case we have developed an analytical model for size and time comparisons of the two encoding schemes. Equations 1 and 2 are for index size and number of scans in Equality encoding and equations 3 and 4 are for size and number of scans in our Multilevel encoding scheme.

Size=C 1)

No. of scans=C/2 2)

Size=$C/n + \log_2 n + d$ 3)

No. of scans = $C/2n + \log_2 n + d$ 4)

The component d represents the size of mapping table. Since we assume that our mapping table will always reside in main memory we consider d to be zero. Figure 5 and Figure 6 show the graph between cardinality verses size for bin sizes 2 and 4 i.e for n=2 and n=4. Figure 7 and Figure 8 shows the graph between cardinality and no. of scans for bin sizes 2 and 4.

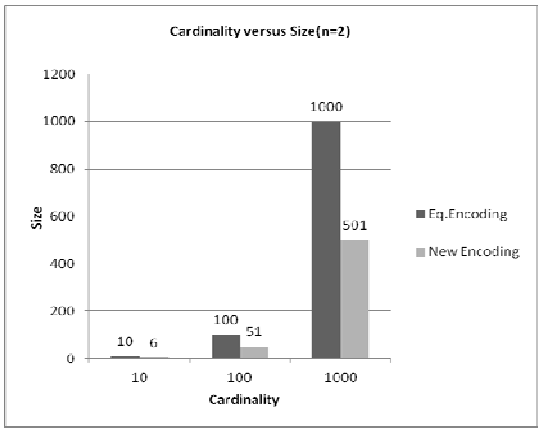


Fig 5. Cardinality vs Size

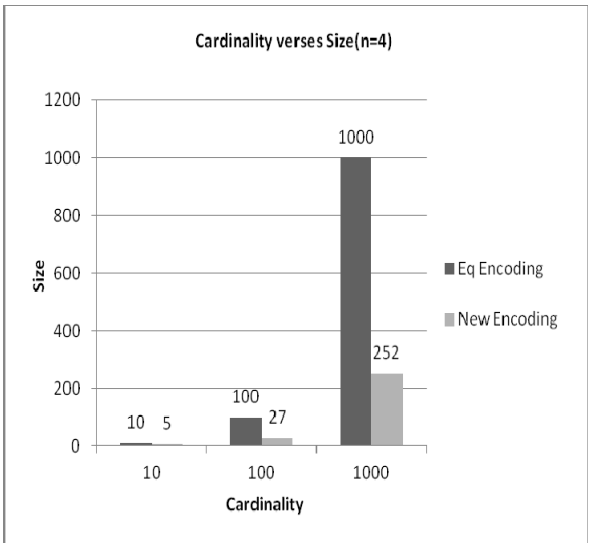


Fig 6. Cardinality vs Size

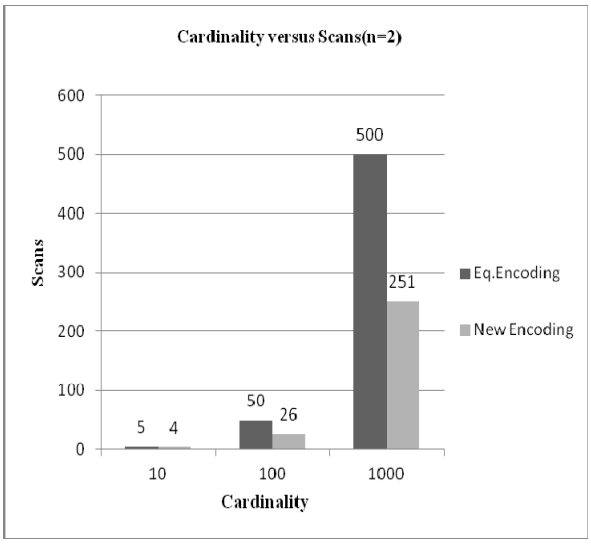


Fig. 7. Cardinality vs Scans

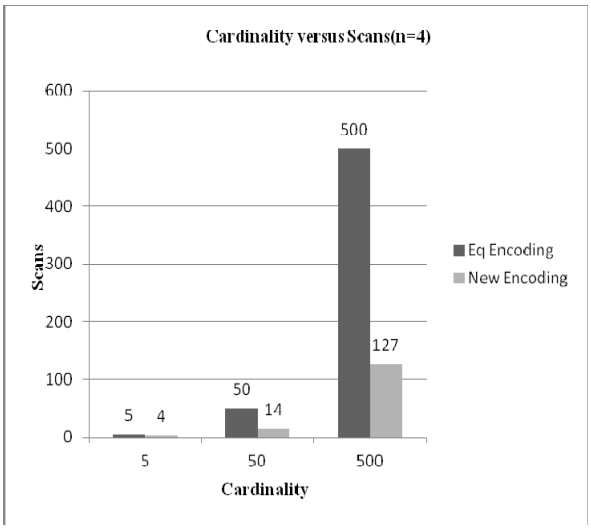
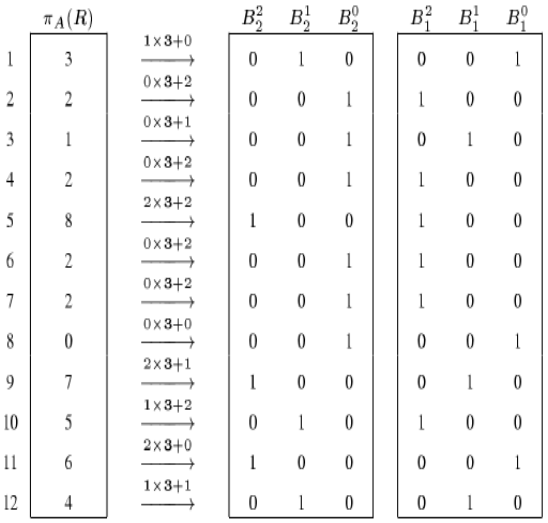


Fig 8. Cardinality vs Scans

5 Multi-component Encoding

The multi-component index are constructed from three basic encoding schemes by decomposing the attribute value into multiple components. The attribute value decomposition defines the arithmetic to represent the values of an attribute. It is the decomposition of an attribute's value in digits according to a chosen base. For example, consider an attribute with cardinality B = 50. An attribute value of 35 can be defined as a single base-50 digit (i.e., $35 = 35_{50}$), or as two base-8 digits (i.e., $35 = 4_8 3_8$), and so on. To make things clear we take the same 12-record relation used in Figure1.



	$\pi_A(R)$		B_2^2	B_2^1	B_2^0	B_1^2	B_1^1	B_1^0
1	3	1×3+0	0	1	0	0	0	1
2	2	0×3+2	0	0	1	1	0	0
3	1	0×3+1	0	0	1	0	1	0
4	2	0×3+2	0	0	1	1	0	0
5	8	2×3+2	1	0	0	1	0	0
6	2	0×3+2	0	0	1	1	0	0
7	2	0×3+2	0	0	1	1	0	0
8	0	0×3+0	0	0	1	0	0	1
9	7	2×3+1	1	0	0	0	1	0
10	5	1×3+2	0	1	0	1	0	0
11	6	2×3+0	1	0	0	0	0	1
12	4	1×3+1	0	1	0	0	1	0

Fig 9. A 2-Component index with base <3,3>

and transform it in a base <3,3> multi-component index .The attribute value 3 can be written in base-3 as $1_3 0_3$. By doing so the bitmaps have been reduced to 6. Figure 9 shows a 2-Component index with base <3,3>. The one-component encoding methods, such as the one used in the basic bitmap index (Figure 2a), requires the largest number of bitmaps. In contrast, the binary encoding produces the least number of bitmaps. This encoding method uses only $\log_2 B$ bitmaps for an attribute with cardinality B. However to answer a range query interval encoding requires accessing only two bitmaps whereas in binary encoding one has to access all the bitmaps. A number of authors have proposed strategies to find the balance between space and time requirements. One main purpose of studying multi-component encoding is to find whether any multi-component encoding can perform better than these two.

- ***Methodology***

Let attribute A have cardinality 1000, let its values range from 0 to 999. These values may be broken into three components of base size 10 each. Each of these components would be a digit of a 3 digit decimal number. Let i_1, i_2, and i_3 denote the values of three components ,the relation among them can be written as $i=i_1+10i_2+100i_3$. Such a three component index can be viewed as composed of three separate indexes on i_1, i_2, and i_3 .We propose BCD encoding for each of these three components. Figure 4 shows an example of our proposed multi-component technique. We observe that for B=0 to B=999 our encoding scheme requires 12 bitmaps. Let us look at the encoding of value 345 which is highlighted in the figure. The least significant four bitmap vectors (B0-B3) are BCD representation of least significant digit of the attribute value.

RID	$\Pi_A(R)$	B11	B10	B9	B8	B7	B6	B5	B4	B3	B2	B1	B0
1	136	0	0	0	1	0	0	1	1	0	1	1	0
2	345	0	0	1	1	0	1	0	0	0	1	0	1
3	789	0	1	1	1	1	0	0	0	1	0	0	1
4	069	0	0	0	0	0	1	1	0	1	0	0	1

Fig 10. Four bitmap vectors

The middle four vectors (B4-B7) are BCD representation of middle digit. The most significant four bitmap vectors BCD

representation of the most significant

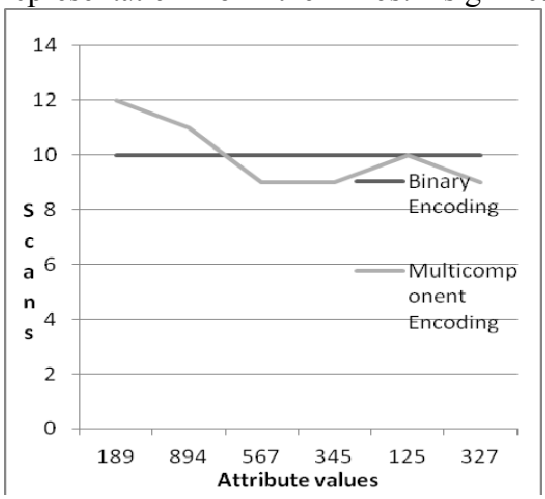


Fig 11. Binary Encoding vs. Multicomponent Encoding

digit of the attribute value. Based on the values of query condition the corresponding bitmap vectors are scanned based on their weights. The retrieval functions f_m for digits between 0 to 9 is given :

$F_0 = \mathbf{B_n}' \cdot \mathbf{B}_{n+1}' \cdot \mathbf{B}_{n+2}' \cdot \mathbf{B}_{n+3}'$;

$F_1 = \mathbf{B_n} \cdot \mathbf{B}_{n+1}' \cdot \mathbf{B}_{n+2}' \cdot \mathbf{B}_{n+3}'$;

$F_2 = \mathbf{B_n}' \cdot \mathbf{B}_{n+1} \cdot \mathbf{B}_{n+2}'$;

$F_3 = \mathbf{B_n}' \cdot \mathbf{B}_{n+1} \cdot \mathbf{B}_{n+2}$;

$F_4 = \mathbf{B_n}' \cdot \mathbf{B}_{n+1}' \cdot \mathbf{B}_{n+2}$;

$F_5 = \mathbf{B_n} \cdot \mathbf{B}_{n+1}' \cdot \mathbf{B}_{n+2}$;

$F_6 = \mathbf{B_n}' \cdot \mathbf{B}_{n+1} \cdot \mathbf{B}_{n+2}$;

$F_7 = \mathbf{B_n} \cdot \mathbf{B}_{n+1} \cdot \mathbf{B}_{n+2}$;

$F_8 = \mathbf{B_n}' \cdot \mathbf{B}_{n+1} \cdot \mathbf{B}_{n+2}' \cdot \mathbf{B}_{n+3}$;

$F_9 = \mathbf{B_n}' \cdot \mathbf{B}_{n+1}' \cdot \mathbf{B}_{n+2}' \cdot \mathbf{B}_{n+3}$;

Where the n can take values 0, 4 and 9.

Consider for instance the Query condition A=345. The multi-component encoding scheme B0, B1 and B2 are scanned for digit 5, B4 B5 and B6 are scanned for digit 4 and B8, B9 and B10 are scanned for digit 3. Thus we see that less bitmap vectors are scanned as compared to all bitmap scans in case of binary encoding.

1. Binary Encoding

Case of equality queries(A=345) :-

TIME:10(B) bitmaps needs to be scanned

SPACE: 10(B) bitmaps are formed.

Case of Range queries(A<=345):-

TIME: 10(B) bitmaps need to be scanned.

SPACE: 10 bitmaps

2. Multi-component encoding:

Case of Equality queries(A=345):-

TIME: 9 bitmap need to be scanned.

SPACE:12 bitmaps are formed.

Case of Range queries(A<=345):-

TIME:9 bitmaps need to be scanned.

SPACE:12 bitmaps.

From the above we analyze that our proposed multi-component indexing technique uses more space than the binary encoded index. However it accesses lesser bitmaps to answer the query Therefore it is possible that that the multi-component index may actually require less I/O time than a binary encoded index. For the given set of attribute values the comparison between no. of scans for different conditions of range queries is given in Figure 11.

A multi-component encoding is usually constructed with some user input parameters. For example if a user chooses the number of components, then it is possible to automatically decide the size of each component to minimize the number of bitmaps generated. For instance ,if a user specified to use a two component encoding for attribute cardinality of 100 ,then each component of size 10 is a good option . An alternative to fixing the number of components is fixing the base size of each component and use as many components as necessary to represent all the attribute values.

- ***Analytical model.***

Let B denote the cardinality of the attribute. For a given attribute value a_i the multi-component encoding decomposes i into a set of integers(i_1, i_2, i_3,…i_k). LetC_1, C_2…C_K denote the sizes of a k-component encoding basis sizes. Using BCD encoding, each component has 4 bitmaps. Thus The total number of bitmaps is D^E=4k .Based on the above assumptions we have developed an

analytical model for size and time comparison of two encoding schemes. Equations 5 and 6 represent index size and number of scans in binary encoding and equations 7 and 8 for size and number of scans in multi-component encoding scheme for range queries

Space= $\log_2 B$ 5)

No. of scans= $\log_2 B$ 6)

Space= 4k 7)

4k>=No. of scans>=k[($\sqrt[k]{B}$)/2 -2] 8)

Thus we conclude that a multi-component index with a base size greater than 2 uses more space than the binary encoded index, however, it may only accesses some of the bitmaps in order to answer a query. Figure 12 and Figure 13 shows the graph between cardinality verses size and the graph between cardinality and number of scans for the binary encoding and our proposed multi-component encoding. The dark shaded portion of bar representing multi-component encoding in Figure 12 tells us about the range over which it may vary compared to Binary encoding. For cardinality B=100,a binary encoding requires seven bitmaps to be scanned whereas our multi-component encoding may require between six(three bitmaps for each digit) and eight(four bitmaps for each digit). Number of scans as per equation 8) will be between 8(4k) and 6

(k[($\sqrt[k]{B}$)/2 -2]). Therefore, it is possible

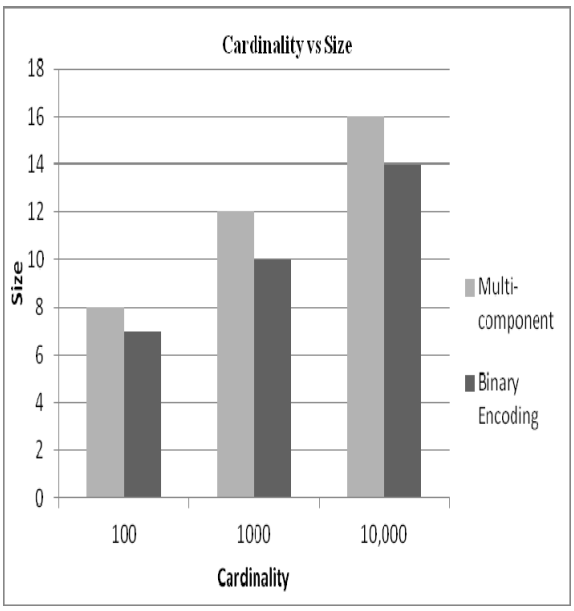


Fig 12. Cardinality vs. Size

that a multi-component index may actually require less I/O time than a Binary encoded index.

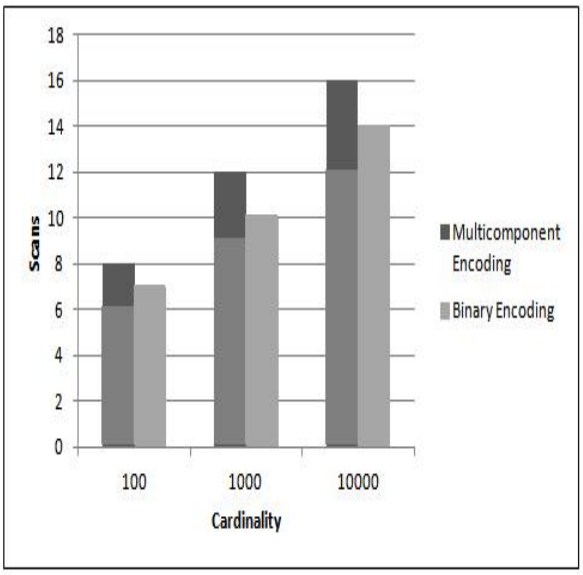


Fig 13. Cardinality vs. Size

Conclusion

The ability to extract data to answer complex, iterative, and ad hoc queries quickly is a critical issue for data warehouse applications. A good indexing technique is reduce I/O intensive table accesses against

large data warehouse tables. The challenge is to find an appropriate index type that would improve the queries' performance. Various Bitmap indexes have been demonstrated to significantly speed up searching operations in Data warehousing and On-Line Analytical Processing. All encoding methods proposed in the past can be categorized as either a Multi-component encoding or Multilevel-encoding. In this paper we have present novel variations in the class of Multi-level and Multi-component indexes and find that they answer range queries faster than some of the existing multi-level and multi-component indexes. We have formed an analytical model to predict the index size and access time of these encoding schemes for worst case scenario. The main contribution of this paper is the development of equations that predict the index size and number of scans which is a measure of I/O operations. Future work may include conducting a experimental evaluation of these two proposed encoding schemes on real application data.

References:

[1] Surajit Chaudhuri and Umeshwar Dayal. An Overview of Data Warehousing and OLAP Technology. In Proc. ACM SIGMOD Conf.1997, Pages 65-74.

[2] C.Y. Chan and Y.E. Ioannidis. An Efficient Bitmap Encoding Scheme for Selection Queries. Computer Sciences Department, University of Wisconsin-Madison,1998. http:/ /www.cs.wisc.edu/Ncychan/interval.ps

[3] Patrick O'Neil and Dallan Quass. Improved-Query-Performance with Variant Indexes. In Proc. ACM SIGMOD Conf. 1997, Pages 38-49.

[4] P. O'Neil and G. Graefe. Multi-Table Joins Through Bitmapped Join Indices. ACM SIGMOD Record, pages 8-11, September 1995.

[5] Morteza Z,Somnuk P.,Su-Cheng Haw. An adequate Design for Large Data Warehouse Systems: Bitmap index versus B-tree index. International Journal of Computers and Communications, Issue 2, Volume 2,2008.

[6] Stockinger, K and Wu, K. 2006. Bitmap Indicies for Data Warehouses. Idea Group, Inc., Chapter VII,179-202.LBNL-59952.

[7] J. Winchell. Rushmore's Bald Spot. *DBMS*, 4(10):58, September 1991.

[8] O'Neil, P. Model 204 Architecture and Performance. Workshop in High Performanc Transaction Systems, Asilomar, California, USA. Springer-Verlag.1987.

[9] Wu, K., & Otoo, E.J., & Shoshani, A. (2004). On the Performance of Bitmap Indices for High Cardinality Attributes. International Conference on VLDB,Toronto, Canada. Morgan Kaufmann.

[10] O'Neil.,E., O'Neil,P., and Wu.K., Bitmap index design choices and their performance impli-cations, Database Engineering and Applications Symposium. IDEAS 2007. 11th International, pp. 72-84.

[11] Kesheng Wu, Ekow J. Otoo, and Arie Shoshani. A performance comparison of bitmap indexes. In CIKM, pages 559–561. ACM, 2001.

[12] Kesheng Wu, Arie Shoshani, Kurt Stockinger. (2010) Analyses of Multi-Component Compressed Bitmap Indexes.ACM Transactions on Database Systems Vol . 35,No.1. Article2.

[13] Kesheng Wu, Kurt Stockinger and Arie Shoshani. Performance of Multi-Level and Multi – Component Compressed Bitmap Indexes. Lawrence Berkeley National Laboratory, June 11, 2007.

[14] Kesheng Wu, Ekow Otoo, and Arie Shoshani. Compressing bitmap indexes for faster search operations. In SSDBM'02, pages 99–108, Edinburgh, Scotland, 2002. LBNL-49627.

[15] Kesheng Wu, Ekow J. Otoo, and Arie Shoshani. Compressed bitmap indices for efficient query processing. Technical Report LBNL-47807, Lawrence Berkeley National Laboratory, Berkeley, CA, 2001.

[16] G. Antoshenkov. Byte-aligned bitmap compression. Technical report, Oracle Corp., 1994

[17] K.-L.Wu and P. Yu. Range-based bitmap indexing for high cardinality attributes with skew. Technical Report RC 20449, IBM Watson Research Division, Yorktown Heights, New York, May 1996

[18] Doron Rotem, Kurt Stockinger, Kesheng Wu. Optimizing I/O Costs of Multi-dimensional Queries using Bitmap Indices. DEXA 2005,220-229.

[19] Doron Rotem, Kurt Stockinger, and KeshengWu. Optimizing candidate check costs for bitmap indices.In CIKM 2005. ACM Press.

[20] Wu.K.,Otoo,E.,and hoshani, A.2006. Optimizing bitmap indices with efficient compression. ACM Trans. Datab. Syst 31,1-38

Grid and Data Analyzing and Security

Fatemeh SHOKRI
B.A Software Computer Engineering, Computer Engineering Department, Mazandaran Institute of Technology, Mazandaran, Iran.
fatemeh.shokri@aol.com

This paper examines the importance of secure structures in the process of analyzing and distributing information with aid of Grid-based technologies. The advent of distributed network has provided many practical opportunities for detecting and recording the time of events, and made efforts to identify the events and solve problems of storing information such as being up-to-date and documented. In this regard, the data distribution systems in a network environment should be accurate. As a consequence, a series of continuous and updated data must be at hand. In this case, Grid is the best answer to use data and resource of organizations by common processing.
Keywords*: Grid Computing, Secure Structure, Common Processing*

1 Introduction

As it stands, there are lots of researches have been done regarding using calculation systems based on computer networks. According to their importance and their effects on different aspects of calculation systems, in this regards several researches have been done. In these studies different dimensions of them were studied such as: Evaluation of Job-Scheduling Strategies for Grid Computing (Hamscher et al, 2000), UNICORE: A Grid Computing Environment (Erwin, 2001), Intrinsic vulnerability assessment of the aquifer in the Riana spring catchment by the method SINTACS (Janza and Prestor, 2002), Economic models for resource management and scheduling in Grid computing (Buyyal, 2003), Grid Computing: A Brief Technology Analysis (Smith, 2004), Trusted Grid Computing with Security Binding and Trust Integration (Song, Hwang and Kwok, 2005), Scheduling Algorithms for Grid Computing: State of the Art and Open Problems (Dong and Akl,2006), Introducing Virtual Private Overlay Network services in large scale Grid infrastructures (Palmieri, 2007), Implementation of Computational Grid Services in Enterprise Grid Environments (Richard, Joshi and Eswaran, 2008), Reliability in grid computing
Systems (Dabrowski, 2009), Reliable Job Scheduler using RFOH in Grid Computing (Mohammad Khanli,Etminan Far and Ghaffari, 2010), Trust Based Authorization Framework for Grid Services (Singh, 2011), A Time-minimization Dynamic Job Grouping-based Scheduling in Grid Computing (Mishra, Mohanty and Mund, 2012).

Due to the expansion of electronic services that are available to share resources in a virtual operating environment, security and trust are important to the parties participating in this environment. Therefore, for identification and understanding of the environment, we should examine the structure of these environments. By expanding various structures such as integrated data environments, data bases virtual networks, Private networks, wide-area networks, and increasing the range of information sharing in a virtual domain, the need to examine and audit the environment in order to gain trust - (either participant) is recommended. If the users and participants

in virtual environments are demanding an independent unit that has sufficient knowledge of the system and can investigate, review this area and participants should be provided with the strengths and weaknesses of safety control structures.

According to the necessity to audit this environment, the auditor's understanding is the first steps to study and review. Required knowledge is understanding of Mentioned environment. Auditor's understanding of the environment in which helps him to deliver efficient and effective investigation. To properly audit each area of the tool, it must be used in accordance with the environment and must be aware of the work environment. The amount of investigations required for each stage depends on the auditor's previous knowledge of the activities and structure and the system of the environment. If the auditor has sufficient and comprehensive knowledge about Environmental audit, the audit procedures would be a more specific, investigation time reduces and reporting speed increases. In this context, this paper examines the structures of the Grid environment and safety grid types, how to establish a secure connection on the grid, the grid safety standards, safety rules on a shared processing environment.

What is the Grid ?

Grid is said for system that is used for the management and integration of distributed resources and services within Domain Controllers. Grid in virtual organization in comparison with private groups, dependent resources and services has a common goal. To meet the needs of information integration and job control management among virtual organizations has been created.**(**Von Welch et al, 2005)

Grid computing or networks connected to the computer, is a new network model that performs massive computing by using attached processors. Grid uses resource of computers that are connected to networks and can do complex computing with resultant force of these resources. They do this by splitting the operation and consigning the piece to the computer in network to do. (Shahcheraghi and Ahmadinia, 2011a)

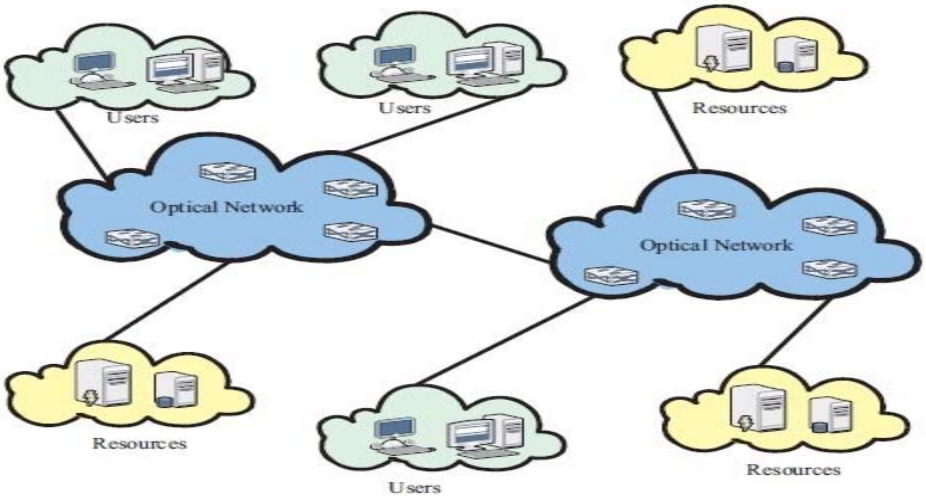


Fig 1. An Example of Grid computing networks

Grid computing networks proposed in the late 1990s as a replacement for conventional supercomputers emerged to address specific problems that require numerical computation and access to large volumes of data that had been distributed.

The main idea was that the networks that are fast enough and using appropriate software, numerous research groups that were

geographically dispersed, could share Computing resources and data management resources in a single system. As result, the system is able to cope with the issues involved with each of these groups could not deal with. (Shahcheraghi and Ahmadinia, 2011c)

Grid computing is a hardware or software structure which provides a range of reliable, stable access to the features and capabilities of off unused or excess of the current needs of the participants in the system. (Foster, 2002)

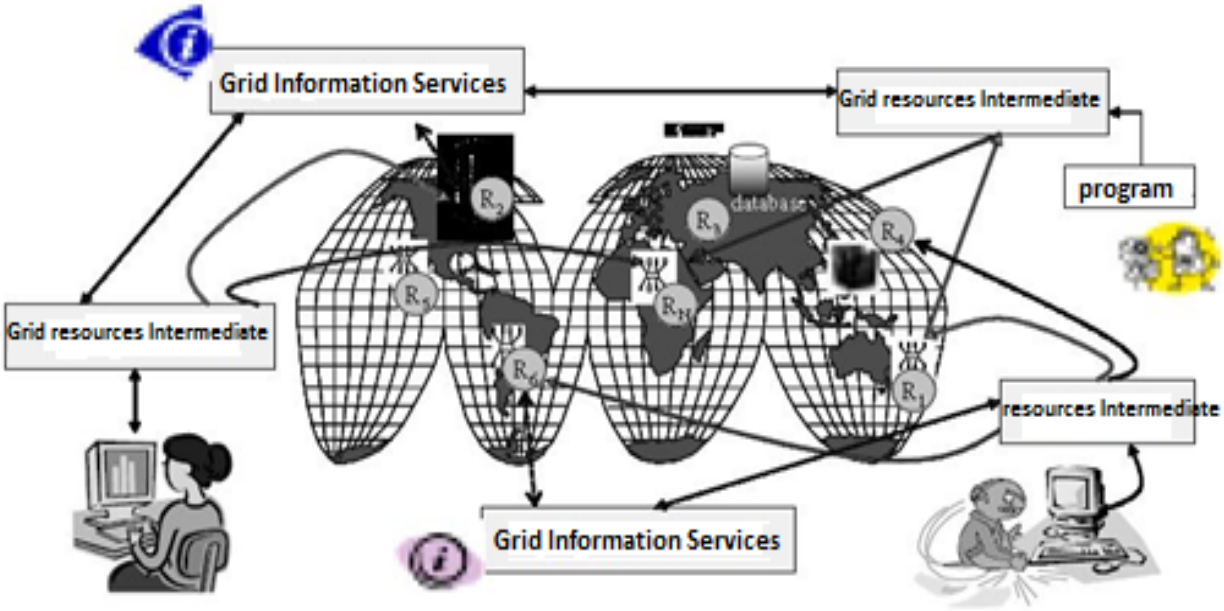


Fig 2. the grid virtualizes geographically disperse resources

In this regard, grid dynamically gathers data Sources from virtual organization (VO) where a service provider that manages and display resource integrally.

In addition, it ensures safety of data in these organizations.

Of a project that can use a Grid infrastructure as follows:

1. **SETI@home** is One of the most popular processing cycles network used 3 million computer to obtain information chain.
2. Another known project is"**distributed.net**" Which began in 1997, 1100 successful projects in distributed information can be found in its records.
3. Advanced Computer Facility at NASA, in the heritability of cognitive algorithms, a processor cycle, which employs the condor, runs nearly 350 workstation of solar and SGL.
4. Up to 27 April 2007, plan to integrate the information in organizations, created based on the mp network and product of processes cycle personal computers is **pcs**, which connected one million the database to each other.
5. The project "**Enabling Grids for E-science**" that is run by the Europe Union, including a connection to science centers in Europe and United States in order to integrate information in scientific projects, use the capacity of useless of the existing scientific resources .Cases which are represented are only a sample of Grid use in the world.

Therefore, the grid uses safety models like GT to create a safe zone. In the following, we examined part of this safety model.

GT2 Security Model for Grid Resource Allocation and Management (GRAM), Monitoring and Understanding (MDS) and is responsible for transferring data from protocol FTP file transfer.

The security model uses the grid security infrastructure (GSI) and the GSI uses three basic elements in its safety mechanism.

The three basic elements should be mutually trusted in formation of a virtual domain

To meet the needs of users in the Grid environment by using Combination of the dynamic style,

Have important roles, as follows:

1. Multiple safety mechanisms for participatory organizations or environments:
   Often it is spoken the important investment in the creation of a mechanism or structure for optimum safety, overshadowed performance some group to be able to create a stronger environment for sharing resources.
2. Creating a dynamic services: Users should create new services (according to sources) dynamically and non-intervention of the controller.

These services should participatory (jointly) and have safe interaction with the other participants. These services need to be done that are not inconsistent with the methods of local control.

3. Creating dynamic and trusted domain: In order to share resources in the virtual domains, should be a multilateral relationship between resources users and participants. The trust must create between members of the participants, It leads to greater trust between participants and biased behavior reduce.
   For the relationships between participants in a virtual organization to be safe,

The relationship of the organization (which forms the foundation of any organization) should be identified and defined. Gt3 is an enhanced version of the GT2 model that allows Programmers and users to work automatically on the Grid. And it has covered the failure and security breaches of previous model.

In line with the standards, two mechanisms and safety policies in main organizations are required:

1. safe technologies based on the use of virtual organizations as a bridge between the participants in the information-sharing environment:
   Research results show that the system uses widely software, In addition to consideration of the mutual benefits, it prevents from side effect and a way-use.
2. Grid security structure:
   Grid security structure has been established to route and support connection of application systems. Definition of part of the extensive services structure and other elements that has contributed to the new competitiveness for achieving safety opportunity, the following is stated. In this regard, safety standards after the information technology like Safety service structure after the integration of GSI and the open grid Services structure OGSA is provided. Using techniques for compiling and expressing as a authorized application way, is expanding.

Safety structure in virtual information environments:

Since the in the virtual information organization, the accumulation and sharing of information would be dealt with. The safety of such a range should be considered as the most important components. Safety Such an environment has secondary components such as authentication,

authorization, and encryption information. In order to better understand the structure of information used in virtual environments,,

According to (1), three areas where safety is established in the environment, are described in the following:

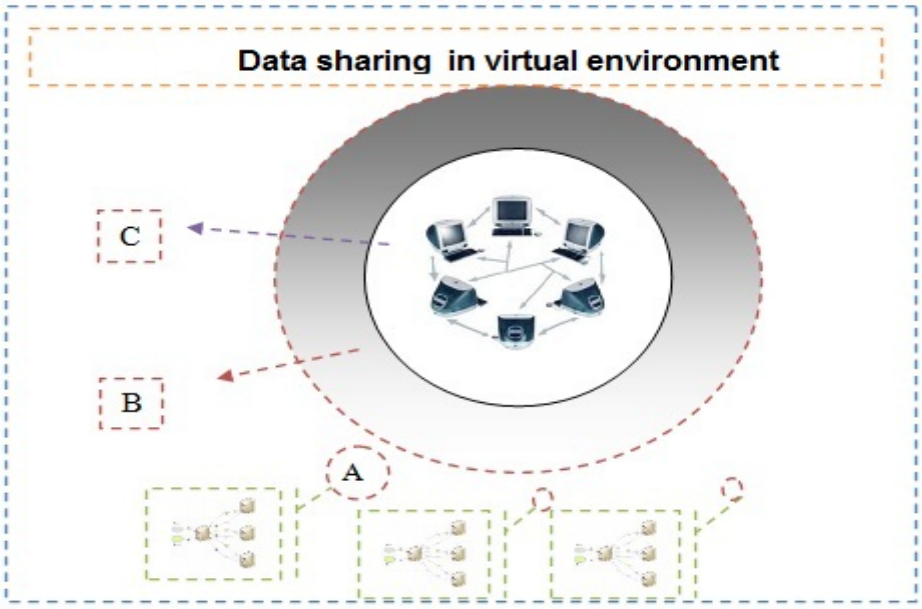


Fig 3. Review of data sharing in virtual environment

1. ection (A): the set of regulations, restrictions and permits to enter the virtual organization (or range) that Participants (prior to entering corporate area) should sign a security agreement to acquire the licenses. In this area investigate the validity and history participating, and How to provide information and resources will be examined.
2. ection B)): In this section the control of hardware and software such as Fire walls, crossing servers, and identification ... (To prevent unauthorized entry or exit of non-immunized) are used It should be noted that in addition to controlling the input and output of the system control of the credibility of the information apply.
3. ection (C), within the range of participation: Host environment provides, the safety agency that have the applicability of the developed environment, This means that determine illegal information with non-optimal levels of performance, some level of participation, How to access the participatory range. However, the safety information structure have specific and exact standard for virtual systems.

   These structures are generally focused on two issues.

   A) y default, insists on establishing communication based on the mutual permissions from the secure lines which

are mutual trust. It put forward The role of the safety certificate which determines crucial condition for communication

B)

ritten access agreement to subject of participation: it provides a simple consensus to achieve goals and use safety protocols such as Web Services Description Language, Multilateral relations with web services that communication between units in different work stations accordance with the specific rules of each service and an exchange of messages, Including items discussed in information safety structure. Some of these mechanisms are described below:

1.

eb Services Description Language (WSDL): It is based on business reporting language that is used to describe Web services. Services as the license from network endpoints or ports are defined, and provide a specific structure for the transcripts according to the respective purpose. This language uses a summary of the constantly and relevant messages to Contact in order to be specified Minimum standards for the use of information in the network.

2.

ommunication Lines: Communication Lines Between subscriber's aspect of validity, Safety and reliability must be examined in terms of proper communication. Communication between the subscribers in this environment (despite the signed agreement) May be Non-secure, unauthorized and misuse resources in the cooperative organization.

Therefore, the grid safety standards have been raised that the two principal fo llows.

- An open grid service architecture (OGSA) this specification in 2002 is suggested by SETI @ home, the following:

A)

y default, establishing communication based on the mutual permissions which are the secure lines and mutual trust. But not ignored the role of certificate and safety lines that provides condition for secure communication.

W

B)

imple Object Access Protocol (SOAP):

Building a simple environment to achieve the objectives of the agreement and the use of safety protocol such the Web Services Description Language, Multilateral relations Web services which provide relation between units in different workstations tailored to the specific rules of each service and an exchange of messages.

Evaluation and audit Participatory range

C

Due to the safety issues that have been studied**,** a reliable control environment, the auditor's opinion will influence the audit test.

System where has High capacity for security and control, Data collection and analysis for the auditor to be more specific; however, investigation of the safety structures such a Participatory environment, the auditor will

needs to follow up its work, In this environment, the auditor is unable to use the traditional methods of auditing, System which is in used, has a structure that is felt, the audit should be quite familiar with the modern technology and Their structures.

Mode of investigation:

To evaluate the structure, audit examines the safety sectors B, A (figure A),

And study Safety agreement in the safety sector A, consult with The expert system and the company's lawyer, Login licenses as An example of how authentication can be used to test participants.

After investigating first safety layer and ensuring the accuracy of the Agreement, Legal gaps and safety in agreement and accuracy of entrance permits to range participation, Auditor investigates hardware and software structures of participation range and control in terms of quality and quantity. After ensuring hardware and software controls providing the audit‘s positive comment, third phase is evaluated.

At this stage, the safety mechanisms within the participatory range, addition to applied policies and procedures, need to make sure of the policies and controls. So, the best way to evaluate these controls is using Special audit tests. Because in this tests, auditor introduces himself as a participant to the system and becomes aware of safety rules in the system, and can study and report the strengths and weaknesses in safety range mechanisms within the organization.

Information subscription license services review:

According to figure 3, the auditor has reviewed the CAS, with the use of embedded systems in the database, mode of communication, access Levels to information and authorization on this service, can monitor and recounts error. Considering in the information participation range and resources, a self-adaptive system can be used to increase security. To evaluate the control method, Auditor uses special methods such as artificial intelligence in the evaluation and investigation of the safety structure.

Figure 3 illustrates the process of receiving, storing, indexing and analyzing data. As it can be seen, users can get result of analyzing data in system via their own software and computers and use them for their purposes.

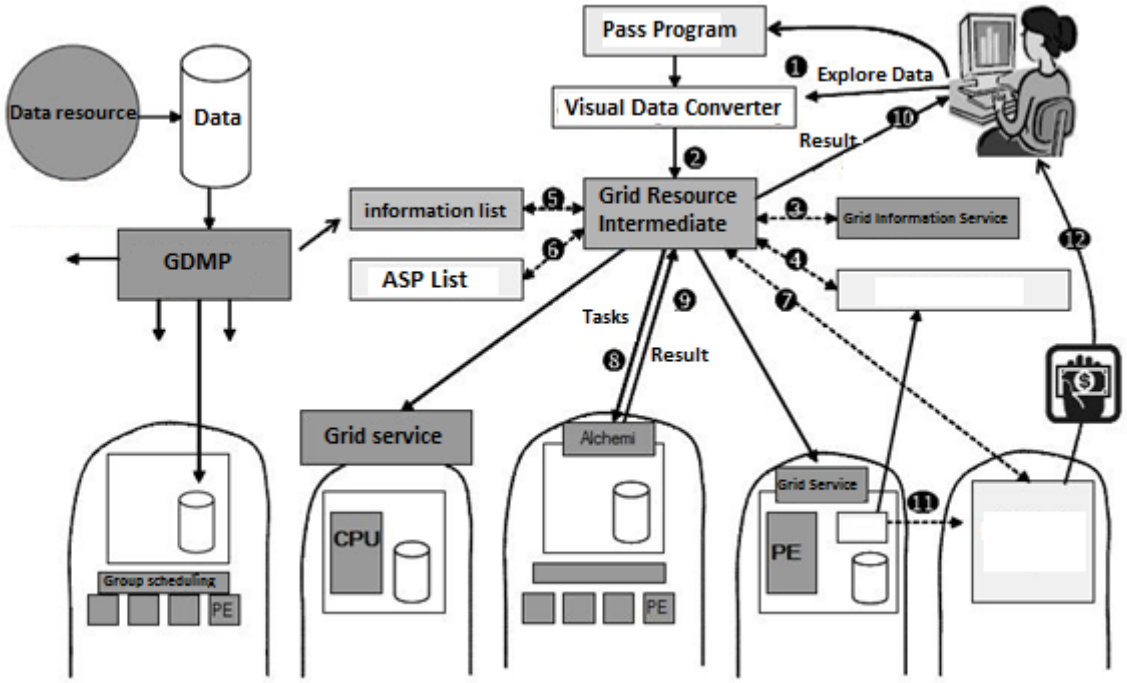


Fig 4. Grid technologies and analyze data distribution

Cooperative computing in grid:

Infrastructure and determining the type of network is data processing methods in computer networks. In other words, the method of processing data in a network is based on computer networks. Network data processing methods are done in three ways: centralized, decentralized, and shared. In Centralized processing, all processing is done at a central computer, and it is suitable for organizations with a centralized structure that requires large computer (in terms of computing and processing power). Decentralized processing is done on personal computers. And instead of using a shared central computer, any user can process information with aid of their personal computer. Difference between centralized and decentralized processing is that the data processing is done in users' computers or terminals. And it creates hardware distribution in organization.(Shahcheraghi and Ahmadinia, 2011b).

In Participatory range of grid systems, processing could be decentralized or distributed, and participants are able to process information jointly (In this environment, a participant can process with use of information from other participants.

This process has the following features:

- ❖ cooperation and resource sharing in processing
- ❖ a continuous network for exchanging sources information between Participants
- ❖ the creation of shared files for cooperative processing
- ❖ a secure environment for processing, without a processing intermediary or handler

Group management Grid system:

The management team make in an important contribution in creating a resource sharing database, Information transfer facilities and systems, developing Local parallel methods of data transmission and combining resources.

This group causes centralized control system that has a complete knowledge of the system states, User demand and exact and comprehensive control on components of participants in the system. This group is fully aware of the protocol goals and safety agreements to support the system. Determining location and time of the participant's access to information in the range is duties of this group. The considerable quality of these services is established in the participatory range.

Conclusions

Grid can be considered as a new experience in virtual research that requires extensive data management. In this article, we review a part of this process. But since Grid is new structures and concepts; for this reason more research is needed in this regard to evaluate more features of this structure. Due to Grid, managers are facing the new opportunities. Therefore, it is important that we continuously assess these systems.

In this regard, I put forward some researches which have done based on Grid technology.

Table 1: A summary of the most important researches on Grid from 2006 up to 2012

Date	Title	Authors	Result
2012	A Time-minimization Dynamic Job Grouping-based Scheduling in Grid Computing	Manoj Kumar Mishra, PrithvirajMohanty, G. B. Mund	This paper proposes "A Time-Minimization Dynamic Grouping-Based Job Scheduling in Grid Computing" with the objective of minimizing overhead time and computation time, thus reducing overall processing time of jobs.
2011	Trust Based Authorization Framework for Grid Services	Sarbjeet Singh	This paper describes different facets associated with trust issues among different entities in a grid environment and proposes a trust model to establish and manage trust relationships
2010	Scientific Data Sharing Using Clustered-Based Data Sharing in Grid Environment	1Rohaya Latip, 2Hamidah Ibrahim and 3Feras Ahmad Al-Hanandeh	In this study, we introduced a new protocol, named Clustered-based Data Sharing (CDS) for data sharing in a large dynamic network such as grid computing by using Clustered-based techniques to improve the accessibility.
2009	Reliability in grid computing systems	Christopher Dabrowski	This study surveys work on grid reliability that has been done in recent years.
2008	A Multi-Agent Architecture		This paper proposed a multi-agent architecture that addressed resource management and application execution with

	for QoS Support in Grid Environment	Ali Rezaee, MasoudRahmani, SaeedParsa, SaharAdabi	support for Quality of Services (QoS) in grid environment.
2007	Introducing Virtual Private Overlay Network services in large scale Grid infrastructures	Francesco Palmieri	In this paper, we propose a novel network resource abstraction for delivering dynamic on-demand Virtual Private Overlay connection services, into large-scale Grid environments.
2006	Implementation of Load Balancing Algorithm in a Grid Computing	1Abdallah Boukerram and 2Samira AitKaciAzzou	This paper describes the complete Implementation of an algorithm of load balancing in an environment of grid computing. The implementation of the algorithm is realized on a cluster of processors in a logic of portability on grids.

References

[1]AlexX.Liu //www.cse.msu.edu/~alexliu2132 , Engineer Building Department of Computer Science and Engineering ,Michigan State University

[2] Christopher Dabrowsk, 2009, Concurrency and Computation: Practice & Experience - A Special Issue from the Open Grid Forum, Volume 21, Issue 8,Pages 927-959, doi:10.1002/cpe.v21:8

[3] Dietmar Erwin, 2001, UNICORE - A Grid Computing Environment, Uniformes Interface für Computing Ressourcen(Final report – in German) http://www.unicore.org.

[4] Fangpeng Dong and Selim G. Akl, 2006, Scheduling Algorithms for Grid Computing:State of the Art and Open Problems, Technical Report No. 2006-504, doi=10.1.1.69.3660

[5] Foster, Ian, What is the Grid? A Three Point Checklist, Argonne National Laboratory & University of Chicago, July 20, 2002

[6] Francesco Palmieri, 2007, Introducing Virtual Private Overlay Network Services in Large Scale Grid Infrastructures, Journal of Computers, Vol 2, No 2 (2007), 61-72, Apr 2007, doi:10.4304/jcp.2.2.61-72

[7] Leyli Mohammad Khanli, Maryam Etminan Far, Ali Ghaffari, 2010, Reliable Job Scheduler Using RFOH in Grid Computing, Journal of Emerging Trends in Computing and Information Sciences, Vol. 1, No. 1, pp43-47.

[8] Manoj Kumar Mishra, PrithvirajMohanty and G B Mund. Article: A Time-Minimization Dynamic Job Grouping-based Scheduling in Grid Computing, 2012,International Journal of Computer Applications 40(16):16-25.doi: 10.5120/5064-7419.

[9] MitjaJANzA&Joerg PRESTOR, 2002, Intrinsic vulnerability assessment of the aquifer in the Ri`ana spring chatcmentby the method SINTACS, GEOLOGIJA 45/2, 401–

406, Ljubljana, doi:10.5474/geologija.2002.039
[10] R.J.A. Richard, Ajay A. Joshi and C. Eswaran, 2008, Implementation of Computational Grid Services in Enterprise Grid Environments, American Journal of Applied Sciences, Volume 5, Issue 11,Pages 1442-1447, DOI: 10.3844/ajassp.2008.1442.1447
[11] RajkumarBuyya, David Abramson, Jonathan Giddy and Heinz Stockinger, 2003, Economic models for resource management and scheduling in Grid computing, Concurrency and Computation: Practice and ExperienceSpecial Issue: Grid Computing Environments, Volume 14, Issue 13-15, pages 1507–1542, DOI: 10.1002/cpe.690
[12] Roger Smith, 2004, Grid Computing: A Brief Technology Analysis, CTOnet.org.
[13] Sarbjeet Singh, 2011, Trust Based Authorization Framework for Grid Services, Journal of Emerging Trends in Computing and Information Sciences, Vol. 2, No. 3, pp136-144.
[14] Shahcheraghi, Azinsadat and Ahmadinia, Hamed, Managing Integration Data in Grid (February 9, 2011). Rayaneh Magazine, Vol. 21, pp. 68-72, 2011. Available at SSRN: http://ssrn.com/abstract=2088074
[15] Shahcheraghi, Azinsadat and Ahmadinia, Hamed, Survey Sharing Information in Dynamic Information Environments - Part 1 (September 23, 2011). Ecommerce and Computer Magazine, Part 1, Vol. 9, No. 59, pp. 45-49, 2011. Available at SSRN: http://ssrn.com/abstract=1968066
[16] Shahcheraghi, Azinsadat and Ahmadinia, Hamed, Survey Sharing Information in Dynamic Information Environments - Part 2 (September 23, 2012). Ecommerce and Computer Magazine, Part 2, Vol. 9, No. 60, pp. 45-49, 2011. Available at SSRN: http://ssrn.com/abstract=2079970
[17] Shanshan Song, Kai Hwang , Yu-kwong Kwok, 2005, Trusted Grid Computing with Security Binding and Trust Integration, Journal of Grid Computing, DOI: 10.1007/s10723-005-5465-x
[18] Volker Hamscher, UweSchwiegelshohn, AchimStreit, RaminYahyapour, 2000, Evaluation of Job-Scheduling Strategies for Grid Computing, Grid Computing — GRID 2000, Lecture Notes in Computer Science Volume 1971, 2000, pp 191-202
[19] Von Welch, Frank Siebenlist, Ian Foster, John Bresnahan, Karl Czajkowski, JarekGawor,Carl Kesselman, Sam Meder, Laura Pearlman , Steven Tuecke, 2005,Security for Grid Services, Universities of Chicago, Southern California & Argonne National Laboratory.

Semi-Distributed Vacuuming Model on Temporal Database (SDVMT)

Mohammad Shabanali FAMI1, Elham Shabanali FAMI2, Dr Mohammad Ali MONTAZERI3, Dr Mohammad Taghi ISAAI4

1Islamic Azad University , emfami @ hrgsoft.com

2Isfahan university of technology, e.shabanali @ ec.iut.ac.ir

3Isfahan university of technology, Montazeri @ cc.iut.ac.ir

4Sharif university of technology, Isaai @ Sharif.edu

Temporal database is one of the most common types of databases. Portfolio management, accounting, storage, treatment management systems, aerology systems and scheduling are applications which their data have time references. Temporal nature of data and increasing size of temporal databases due to non-removal data requires presenting a solution to overcome this limitation. In this research, firstly the current model of vacuuming systems are simulated and analyzed. Then the proposed model introduced for vacuuming systems using distribution concepts. This model is simulated in the same conditions with current model. Using experimental results, advantages and disadvantages of both models were investigated. The proposed model is more capable than the current model in answering temporal queries. Its response time to temporal queries is less than the current model. But the proposed model's cost is more than the current model. Considering the possibility of idle resources usage in organizations, these costs can be ignored along with optimize usage of facilities.

Keywords*: vacuuming, Multiclass queuing model, Schema versioning, temporal database.*

1 Introduction

Temporal database is one of the most common types of databases that its data have time references [1]. Among many different applications of these databases, Portfolio management systems, accounting, banking, aerology systems and scheduling can be mentioned [1]. In temporal databases in contrast with other databases, data will never remove from database. It means that temporal database uses append-only policy instead of update-in-place policy of other databases [1].

Due to increasing size of these databases, introducing a solution for overcoming growing volume of database is required. To deal with this problem, database designers had presented many models such as vacuuming data, schema versioning and aging data management.

Jensen introduced temporal data vacuuming [9]. Skyt studied data management methods for physically removed data [10] and suggested a framework for vacuuming temporal data [2]. Roddick aim was preventing some relations from removing in vacuuming process, so he searched about schema versioning [11] [6]. He also did researches about data mining on temporal

database systems [3]. Jensen presented a framework for vacuuming temporal data. In this framework he vacuumed data base on organization's rules [4]. Grandi studied schema versioning on object oriented databases [5]. Skyt presented a method for removing data based on their features [12]. In all of these methods, some parts of data have been physically removed from database and it partitioned data in active and

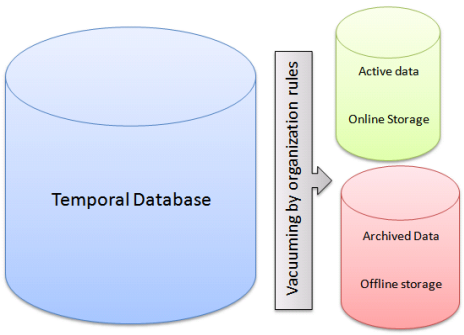


Fig. **1**: the current method for dealing with increasing volume of temporal database. To control temporal database volume, data will be archived base on organization rules.

inactive parts. Inactive data maintained in lateral storage devices, while active data will be remained in online system. In these systems inactive data will removed from online system physically.

To simulate and analyze the current method, a multiclass queuing model is introduced for current vacuuming method. After simulating current model, the proposed method is suggested and simulated by introducing its model. The assessment of proposed model is done by fixing both models' parameters and varying the amount of temporal queries with respect to ordinary queries.

In this paper, the amount of responses to queries and response time parameters was evaluated and considerable amount of growth in answering temporal queries was determined. These observations also show that the response time of temporal queries occurred in shorter time interval than current model. Due to existence of idle resources in organizations, the proposed method makes optimum usage of facilities possible.

In the rest of this paper, firstly the presented model for current vacuuming data methods explained in section 2. Then in section 3 the proposed model for vacuuming data introduced. These models are simulated in the same circumstances and the results obtained are shown in section 4. At last final conclusion of whole paper was states in section 5.

2. Current Vacuuming Methods

In this paper, at first current methods was investigated and it was found that most of them use the same logic. Some of these methods applied data vacuuming to confront infinite increase in database volume. Some others used schema versioning for this purpose. Also aging data management was presented by some others. All of these methods used part-of-data-deletion logic due to conditions.

In this logic, some parts of system's data will be removed base on organization rules. These removed data is maintained in lateral storage devices inactively. Fig. 1 shows current vacuuming systems procedure.

Considering the presented vacuuming method, the model in Fig. 2 visualizes a model for current systems. This is a two-class queuing model which two types of queries enter into queues. These types of queries are temporal queries and ordinary queries. Temporal queries rely on fetching data from inactive vacuums, while ordinary queries will be answered by online system directly.

In model of Fig. 2, D is online server queue. This server directly answers ordinary queries. Activation of inactive vacuums is

undertaken by queue M. Since this routine is non-automatic, it often needs more time than automatic systems to retrieve data. To be able in answering temporal queries, activation of required vacuum should be done. When a vacuum retrieved, it may require another vacuum by itself and it will be continued for some number of vacuums. It means that when a vacuum activated and received its required service from online server, it may enter to queue M for retrieving vacuums that it needs.

If we consider average response time of online queries equal to one second, retrieval time of inactive data and activation time of them will be several times of it. If system administrator will finds a vacuum and activate it in 20 minutes, its speed is about 1000 times less than system's retrieval. It should be noted that there is not an ideal condition whole the time, sometimes the time between making a request for finding correspond vacuums and activation of them will be more than this amount of time. So it is rational to assume queue M slower than queue D for 1000 times.

In Fig. 2, there are two classes of ordinary queries and temporal queries. When the system starts up, the amount of data is less than the power of server, so there is no vacuuming and no temporal query. As time passed, the volume of data will be increased and vacuums will be created.

Fig. 3 that is presented by Ponniah [7]

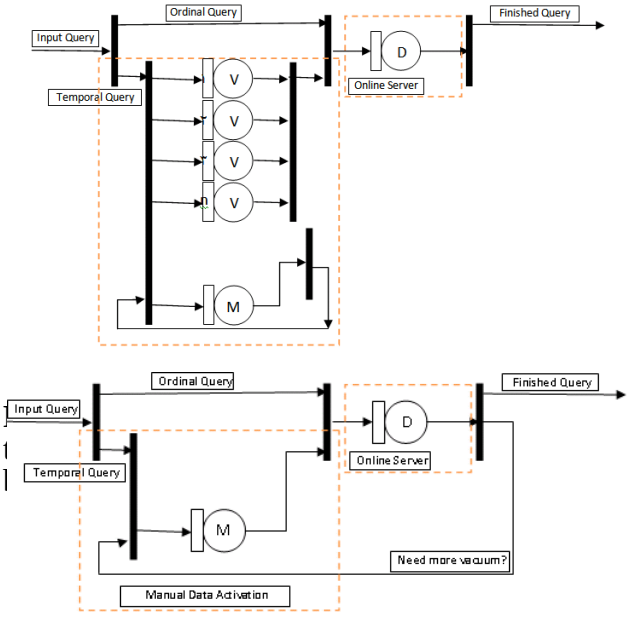


Fig. 2: Modelling of current vacuuming method. In this model, requested vacuums will be activated in queue M

shows that demand for information and number of data systems grow linearly with respect to the time [1]. Suppose that the volume of database at startup time is V_0 and growth rate of data can be calculated according to V_r/T_r. So the volume of database at time t can be obtained by Eq. 1:

$$V = \frac{V_r}{T_r} t + v_0 \qquad \text{Eq. 1}$$

3. Proposed Method

The most important problems of the current model were incapability for answering most of temporal queries and its high response time for other temporal queries. The proposed model that is showed in Fig. 4 has been designed to solve these problems with the objective of optimum utilization of resources. In this model, the concepts of distributed systems are used to improve on pre-designed models. In this system, the temporal vacuums data rather than being kept in inactive storage resources will be kept in on-line servers. Since most organizations usually provide the appropriate hardware infrastructure that does not use optimally, presenting this model provides a method for using maximum power of resources to troubleshoot problems about inactive data.

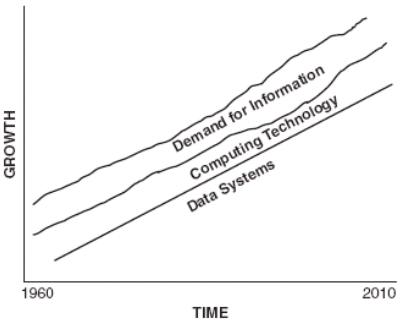


Fig. 3: the growth rate of demand for information with respect to time [7]. As time passed, demand for information and number of data systems will increase

The main difference between proposed method and current method is in optimum usage of organization resources to deliver better services to applicants. More resources possession of the proposed method and scalability of it that obtained from its distributed nature make higher

accountability for this method. If required resources of proposed method were not provided, organization has to use current method. In this situation, however some part of data will be kept inactive, there are more resources to return vacuums and maintain them online for organization.

As an instance, consider a small hospital that it has 20 workstations with normal capabilities along with its online server. This hospital can use its workstations as servers for vacuums. These workstations always have some amount of computational capacity and free storages that can be used for storing and retrieving vacuums data. It is obvious that there are limitations on these resources and after a while the organization will need inactive storage. By the way by optimum usage of resources that was costly for organization, the severity of the problem and the number of inactive vacuums will reduce.

In the proposed model, rather than lateral storage devices, data will be stored actively in some servers called vacuum servers. Vacuum servers are always slower and weaker than online server. When a temporal query arrived, it will be sent to related vacuum server. Then online server will gather and combine all results and answer applicant. In this method, vacuum servers will search for user answer simultaneously.

4. Experimental Results

With the aim of making a comparison between the proposed and current model, both of them was simulated using Simulink part of Matlab R2009a. To have a fair comparison, the adjustable parameters of both models assumed to be similar.

In these models, V_{cur} is a parameter that shows the volume of online database. Similarly V_{vac} is another parameter of the model that indicates average volume of each vacuum. T is the percentage of temporal queries than ordinary queries, V_0 is the primary volume of database and V_r/T_r is the growth rate of the database in time. So there are parameters $V_0, V_r, T_r, V_{cur}, V_{vac}$ and T for simulating models. In both models, primary volume of database is 4 megabytes, online database volume is 20 gigabytes, growth rate of data is 6.6 megabytes in time and average volume of each vacuum is 4 gigabytes.

Simulation was ran for 10, 50, 90 and 99 percent ratios of temporal queries than ordinary queries for 7000 time slices. Using central limit theorem [8] and considering the abounded amount of queries, the service time of online server for both models assumed to have a normal distribution which mean equals to 0.005 and standard deviation equals to 0.0001.

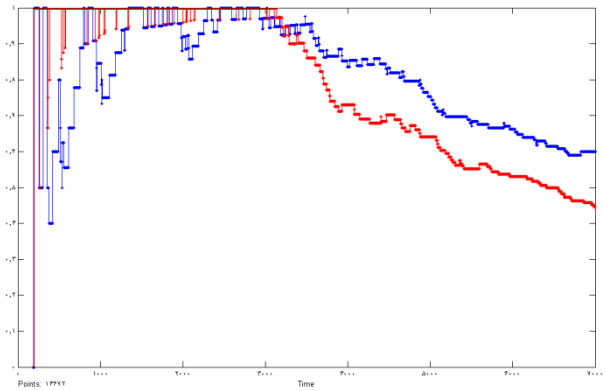
User departure time of current model was considered 4000 time slice. For the proposed model, this parameter was 2000 time slice. Simulation results show that if user patience will be assumed similar in both models, almost all users of current model will be gave up. Consequently user patience of current model is considered more.

4.1 Response Rate per Queries

Obtained results of simulation of both models are shown in TABLE I. In this table, variable T is the percentage of temporal queries than ordinal queries. %T is the percentage of response to temporal queries, while %O is the percentage of response to ordinary queries. TO indicates the number of departure queries.

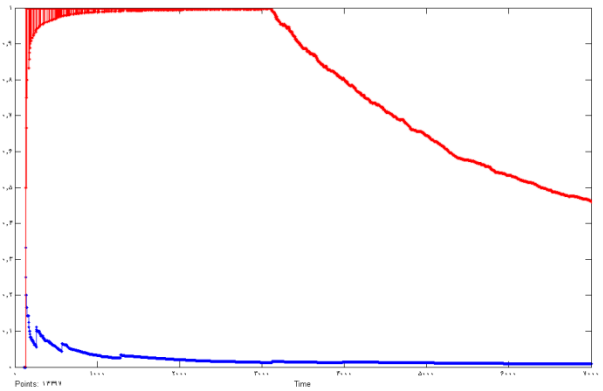
As it can be seen from Table 1, the current system at the best condition can answer 60 percents of temporal queries. When temporal query rate increase, this amount will tend to zero percent. Also when the rate of temporal query increased, the number of departure queries will increase. As it shows in Table 1, the proposed model has higher response power and it is because of parallel usage of vacuum servers and having automatic behavior in contrast with the current model. As temporal query rate increased, response power reduction of model is inevitable.

It can be observed from obtained results of proposed model that systems will answer more queries than current model before occurrence of user departure. About



a

a) When the proposed model has 30 vacuum servers and the ratio of temporal queries is 1%.



b

b) When the proposed model has 30 vacuum servers and the ratio of temporal queries is 10%.

Fig. 5: response rate to temporal queries with respect to time. In these diagrams, red lines are corresponding to current model as well as blue lines are corresponding to proposed model.

ordinary queries both models will answer to whole queries.

4.2 Response Time

In both models when a query is produced, it will be labeled with a time tag. When this query got served entirely, its presence time in the system was calculated and the results are shown in Table 2. Because of parallel processing and automatic behavior of the proposed model, its waiting time reduced considerably. Less waiting time in answering temporal queries, less user give up from getting queries responses.

Response times of ordinary queries in both models are relatively similar, while temporal query's response time in the proposed model is so less than current model.

Table 1: a comparison between response rates temporal queries in percent. Proposed moc will answer more temporal queries than t current model

Proposed model				Current model			
T	%T	%O	TO	T	%T	%O	TC
1	44.33	100	21	1	60	100	0
10	45.97	100	180	10	0.84	100	29
50	43.95	100	939	50	0.14	100	14
90	43.65	100	1711	90	0.35	100	23

Table 2: query's response time. The proposed model reduced the response time of temporal queries.

Proposed model				Current model		
T	Tres	Ores	TO	T	Tres	Ores
1	198.2	0.005	0	1	275	0.005
10	40.01	0.005	180	10	901	0.005
50	17.85	0.005	939	50	311	0.005
90	18.43	0.005	1711	90	605	0.005

4.3 The Number of Completed Temporal Queries

Temporal queries with different rates of 1, 10, 50 and 90 percent of all queries were fed to both models and obtained results are shown in Fig. 6. While the current model makes better results than proposed model for low number of temporal queries, the proposed model will produce better results when temporal query rate increased.

4.4 Spending Time to Service Temporal Queries

In these models, in addition to average response time to temporal queries the number of completed temporal queries in a significant factor. To have more precise comparison, average response time to temporal queries and number of completed temporal queries was measured for both models. The factor of service time of temporal queries was defined by multiplying number of completed temporal queries with average response time of temporal queries.

$$Spending\ time = \frac{}{Response\ time * Number\ of\ completed\ requests}$$

This factor compares both models by considering average response time along with number of responded queries. As it is shown in Fig. 7, proposed model delivers service to temporal queries for more time distances than current model.

5. Conclusion

In this paper the current model for vacuuming temporal data was introduced and simulated. Then to resolve some of its drawbacks a new model was suggested and simulated. At last by determining similar values for parameters of models, response time and response rate of them were compared. Simulation results analyzed and the same behavior on ordinary queries observed. About temporal queries, current model can answer about 24 percent of queries in the best condition. In addition, it produces responses in a long time that applicants will give up their requests. Proposed model answers more temporal queries in a time less than the current model. So the proposed model can be used as a key solution for vacuuming temporal data. One of the most important drawbacks of the proposed model is its implementation cost. Considering idle resources of organization, it can be expressed that this model will make an optimum usage of wasted costs of organization.

Current model has only two servers, active and inactive. If efficiency was quick response to user, when an ordinary query enters to system, since its service will be provided by online server, its response time will be short. Consequently this system has a good efficiency for ordinary queries. In situations that temporal queries are raised, system needs references to inactive database. Besides the response time of inactive servers is high, the efficiency in this situation is not well.

Proposed method needs a short service time to answer ordinary queries by online server and it becomes its efficiency to be high.

About temporal queries two situations may be happened. In some situation organization's resources will provide vacuum servers for system. As a result vacuum servers are responsible to answer temporal queries. As it is known, service time of vacuum servers approximately is equal to online server. Answering temporal queries in this situation needs a short service time and its efficiency is pretty good.

But in situations that organization's resources are unable to provide enough vacuum servers, using current method is inevitable. Consequently some parts of data should be searched in inactive servers. In this situation, system needs higher service time. Since vacuum servers are responsible for some part of queries, this time is less than current model. So the efficiency of this situation is not good but it is higher than the current method yet. In these systems some part of queries will be answered by using vacuum servers and some part of them will be answered by using inactive servers.

To conclude the discussion, proposed model and current model have approximately similar behavior about ordinary queries. Current model has low efficiency about temporal queries. Till

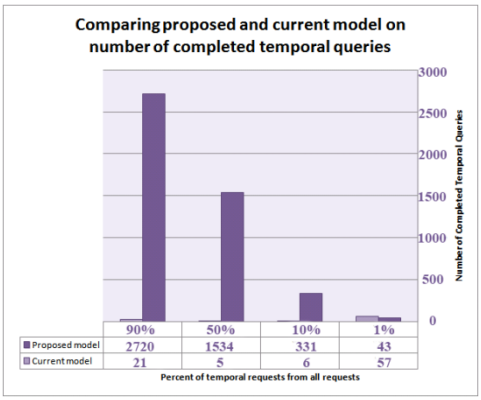


Fig. 6: A comparison between the number of completed temporal queries in proposed model and current model. Histogram bins with light colour are corresponded to current model.

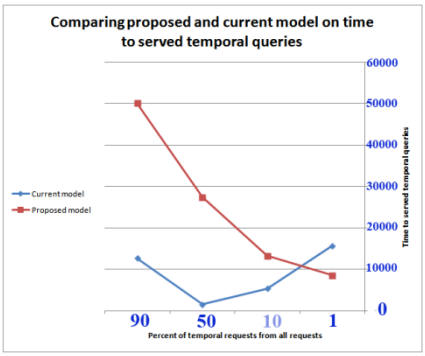


Fig. 7. A comparison between service times of temporal queries for both models. Proposed model spends more times to answer temporal queries than current model.

proposed method has enough vacuum servers its efficiency is good but once they will not sufficient, its efficiency will reduce but it is higher than the current model yet.

6. Future Work

In presenting this method, the performance of all vacuum servers assumed to be equal. Also the importance of all the system's data is the same. In future researches, using intelligent algorithm, a model can be suggested that considers data importance and servers performance. This model will keep important data in the servers with higher performance. These improvements make enhancement in system efficiency because important data need less service time.

One of the solutions to deal with lack of sufficient resources for proposed method is maintaining more widely used vacuums in vacuum servers. These methods previously were used by operating systems when their allocating algorithms keep higher-requested pages actively and lower-requested ones inactively.

References

Periodicals:

[1] C. S. Jensen, "Temporal Data Management" *IEEE Trans. Knowledge and Data engineering*, Vol. 11, No. 1, pp. 36-44, 1999.

[2] J. Skyt and C. S. Jensen and L. Mark, "A foundation for vacuuming temporal databases" *ELSEVIER, Data & Knowledge engineering*, Vol. 44, No. 1, pp. 1-29, 2003.

[3] J. F. Roddick and M. Spiliopoulou, "A Survey of Temporal Knowledge Discovery Paradigms and Methods" *IEEE Trans. Knowledge and Data engineering*, Vol. 14, No. 4, pp. 750-767, 2002

[4] C.S. Jensen and L. Mark, "*A Framework for Vacuuming Temporal Databases*," Technical Report CS-TR-2516, Univ. of Maryland, College Park, 1990

[5] F. Grandi and F. Mandreoli, "*A Formal Model for Temporal Schema Versioning in Object-Oriented Databases*," *ELSEVIER, Data & Knowledge engineering*, Vol. 46, No. 2, pp. 123-167, 2003.

[6] J. F. Roddick, "Schema Vacuuming in Temporal Databases" *IEEE Trans. Knowledge and Data engineering*, Vol. 21, No. 5, pp. 744-747, 2009.

Books:

[7] P. Ponniah, *Database Design and Development: An Essential Guide for IT Professionals*. John Wiley & Sons, pp. 3-10, 2003

[8] H. Fischer, *A History of the Central Limit Theorem: From Classical to Modern Probability Theory*. Springer, pp. 1-10, 2011

Papers from Conference Proceedings (Published):

[9] C. S. Jensen, "Vacuuming," *In Proc. The TSQL2 Temporal Query Language*, Kluwer, pp. 447-460, 1995

[10] J. Skyt and C. S. Jensen, "Managing Aging Data Using Persistent Views," *In Proc. Int. Conf. on Cooperative Information Systems*, Israel, pp. 132-137, 2000

[11] J. F. Roddick, "Schema Versioning," *In Proc. The TSQL2 Temporal Query Language*, Kluwer, pp. 425-446, 1995

[12] J. Skyt, C.S. Jensen, and T.B. Pedersen, "*Specification-Based Data Reduction in Dimensional Data Warehouses*," *In Proc. Int. Conf. on Data Eng*, USA, p. 278, 2002.