Instituto Cultural Minerva

Institute of Brazilian Issues

The George Washington University

Washington, DC

BUILDING AN INFORMATION SYSTEM FOR MANAGERIAL PURPOSES IN THE FINANCE SECRETARIAT OF SAO PAULO STATE

By Paulo de Tarso Leme

Fall 1997

Index

1. Introduction

1.1 Objectives

This paper will study the actual situation of managing information in the Finance Secretariat of Sao Paulo Government, and will then propose data warehouse as a new way to store information.

To understand data warehouse a description of some concepts and the necessary steps for implementing it will be given. Then a rough model of the first data warehouse to be implemented, on tax information, will be given.

After presenting the concept of the data warehouse, a brief description of data mining will be given. It is impossible to implement data mining before the data warehouse. The two are closely connected, but the breadth of this paper will not allow for a detailed discussion of data mining.

1.2 Actual situation in the Finance Secretariat of Sao Paulo

In the Finance Secretariat of Sao Paulo, there are two major areas: Tax Administration Department (CAT - Coordenacao da Administracao Tributaria) and Finance Administration Department (CAF - Coordenacao da Administracao Financeira). This essay will only study the CAT. Below is a description of this department's role, its organization, and its situation in relation to the information system.

1.2.1 Role of CAT

Briefly, this department is responsible for collecting and inspecting three different taxes:

ICMS (Imposto sobre circulacao de mercadorias e servicos) - a tax based on the French model of the Value Added Tax. This tax is collected on every trade operation realized by every member of the productive system.

IPVA (Imposto sobre propriedade de veiculos automotores) - a tax based on the property of automobiles.

ITBI (Imposto sobre transmissao de bens imoveis) - a tax based on the transmission of property of real estate.

1.2.2 Organization

This department is headed by a coordinator and is divided into eighteen regional offices, delegateship (DRT), with each headed by a delegate who supervises all field operations. The regions are divided into 81 districts (IF), administered by inspectors. The districts are divided into 280 centers (PF) led by chiefs. The district offices collect taxes, ascertain delinquent and additional tax liability, investigate violations of internal revenue laws, and aid the public in calculating taxes.

1.2.3 Operational Information System

There is one system named SAFT (Sistema de arrecadacao e fiscalizacao tributaria - Tax Collection and Inspection System) which processes the operational necessities of the department. In this system there are two main documents:

· GIA (Guia de Informacao e apuracao - Information and Calculation Form) In this document the taxpayer reports the value of his purchases, the income resulting from sales, and calculates the tax that he owes to the government. This report is rendered monthly and is sent to the government on a magnetic media. The due tax is the tax of the sales minus the tax already paid for the purchases.(Value added tax).

· GARE (Guia de arrecadacao - Collecting Form) In this document the taxpayer pays the tax owed to the bank and informs the government what kind of tax is being paid.

1.2.4 Executive Information System (EIS)

The operational system SAFT does not provide managerial information at all. Hence, it is necessary once a month to replicate the information from the above documents through a software named SAS. This software allows for some analysis, but it also has some problems:

· Requires complex commands to extract information

· Has no windows interface

· Is not a relational data base

· Does not run on the computer network. Thus, it necessities a permanent line of communication with the mainframe, and it also uses the expensive mainframe computer.

· No easy communication between this data base and text editors (like MS Word) and spreadsheets (like MS Excel).

· Takes a long time to send the necessary information to the managers. Because of the bureaucracy, several official letters are needed to get the information. Usually it takes more then ten days to provide the necessary information. When it arrives, it may be wrong or is no longer needed.

1.3 Information as necessary for the administration and the society

It is not an exaggeration when we say that there is no democracy without information. The shortage of information about our society impairs the organization of it. Information is the basis for correct decision making. It is impossible for our society to evolve if we do not know what and how big the problems are, and what resources are available to solve them.

This paper is about a public organization; therefore, in this kind of organization it is mandatory to make all types of information available to the society. By doing so, the society can know how taxes are collected and spent,

how heavy the burden of the government is, what should be done to improve the efficiency of the government, etc. Knowing these things is part of being a citizen.

Information is important for the operational levels of the company; but processing this type of information is already solved. This is stagnant information. Once it has been provided, it is not necessary to analyze it, it will always be the same.

Managerial information, however, is different. It is always changing according to the volatile business market of modern life. In this case, information is not enough. We need tools to make it possible for the manager (or his assistant) to produce a report (maybe on the screen) containing the necessary information grouped and classified according to the necessities of the moment. It may be information about statistical variables, tendencies, regression, simulations, neural analysis or whatever else is necessary. We must provide the basic information and the tools necessary to manipulate it whenever and however it is necessary.

The process of creating a solution should be fast, although it may take several attempts to produce a satisfactory result. In this case, the process must be very easy and friendly, and the system must be able to manipulate information quickly.

The classical solution of producing a system does not solve this problem. The process of studying what is needed, producing a project, programming and implementing the solution is inappropriate, because when we start the process the problem is unknown. It will be known only when the problem appears. A new strategy must be applied to make the information available and to provide a tool that enables the manager (or his staff) to do it by him/herself.

1.4 Comparison with the situation in other organizations

"Do not reinvent the wheel": this is a very common saying that shows us how important it is to look for other organizations which are facing the same situation and the same difficulties. Therefore, before starting to solve the problem, it is very important to know how similar organizations are dealing with it. Moreover, it is very helpful to read about the problem. To do this we contacted the following organizations: the Federal Revenue of Brazil, some banks, and some industries. It was very reassuring when we realized that there was a high similarity between our reality and the reality of these organizations. Most organizations in Brazil and the world have a similar problem.

We share the following aspects:

· There is an operational system, run on a mainframe computer, which attends to the basic necessities of the operational process.

· There is a lot of information in this operational system. Unfortunately this information is not available for managerial proposals.

· There is a common decision to create a different data base named data warehouse to store all managerial information about the company for a long period of time (at least 5 years).

· This new data base will not be created on the mainframe computer; instead it will be created on the computer type RISC, and will use an operational system following the standard UNIX.

· It is common sense that this process will take a long time. Also, it has to be done step by step, but according to a global project.

2. Proposal for action

2.1 Basic proposal (OLTP, OLAP, Data Warehousing, data mining, RDBMS)

The basic proposal is the implementation of the data warehouse with tools for an analytical process (OLAP) using a Relational Data Base Management System (RDBMS). This will be followed by implementing additional tools for more accurate analysis using data mining.

The following gives a more detailed explanation about these components of the solution.

OLTP (On Line Transaction Process) is a common computer system that has the basic function of controlling the operational procedures of the company.

This system holds a large quantity of information from every operation of the company. Despite the fact that this information is important for the daily operations of the company, it is not the information needed for making decisions.

An operational system has these characteristics:

· Information stored for a short period of time (usually months).

· The process involves only a few records of information. (Usually one register is processed in each operation)

· The operation has to be really fast, because there is usually a customer or a taxpayer waiting for the result. It cannot be delayed more than one minute.

· The process is completely defined with the desired result well known before the development of the system.

· There are few changes in the system after it has been developed. If it satisfies the necessities it is considered finished.

OLAP (On Line Analytical Process) has the purpose of making available all information necessary for the decision making process. It is not interested in the daily operations of the company. Instead, it works to provide information for the correct tactical and strategic management of the organization.

An analytical system has these characteristics:

· Information stored for a long period of time (usually more than five years).

· The process covers many records of information. For one query, it may cover millions of records.

· The operation does not have to be fast, because it will be done only one or two times, and there is no client or taxpayer waiting for the results. It may take more than one hour to be processed, but time is not a real problem.

· In this kind of process there is not a precise question; it is rather an interactive process. After one question (query), there is immediately another one. It is a creative process for we have to discover the logic that there is in the data. What relations exist in it.

· According to the changes in the market, the system has to be changed because there are new necessities, new facts to be analyzed. The system has to record different types of information, because over a long period of time the structure of the information changes many times. The market, nowadays, changes completely in a short time. The velocity of this changes is really bigger than it used to be few years ago.

Data warehouse is a container where the data of an organization is stored. It is the basis for the OLAP. If the data is properly stored then the analytical process will be better, and easier to conduct. A large quantity of

information does not guarantee the ability to make a correct analysis. What is important for analysis is how this information is structured.

Data warehousing refers to the process of how data is stored. It does not consist of just one moment. This process will last for the whole life of the organization. There are some rules that have to be respected during this process. These will be explained during this paper.

Data mining is a set of tools that may be used to make a correct analysis of the data. It is impossible to use data mining tools without the proper data warehouse implemented. This paper does not focus on this in a detailed manner, because what we want to establish is the importance of the data warehouse. However, a brief discussion of the data mining will allow this importance to be seen.

RDBMS (Relational Data Base Management System) is a software responsible for controlling all information that is stored in the Data Base. These responsibilities include:

· Controlling who has the rights to consult or update information.

· Registering the information about the data. It contains the metadata about every data, i.e., the explanation about what the data means, who is responsible for updating this information.

· Creating a log that tell us about who and when the information was updated.

· Identifying the users, checking user codes and passwords.

· Guaranteeing the integrity of the information, validating each field of the record and the relational integrity between the registers. For instance, the data base can not accept a payment for a taxpayer that does not exist.

· Making possible access to the data base from different kinds of software, such as text editors, spreadsheets, graphics etc.

The RDBMS is the basis for both the OLAP and OLTP systems. Without RDBMS there could not be computer science.

2.2 Centralization versus decentralization

This is an important topic to be discussed. Should the information be centralized at just one point or should it be distributed at several points?

At the beginning of the computer science age, data bases were usually centralized because of technical aspects like:

· the equipment was very expensive, and it was not economically viable to buy several computers for each branch or department of the company.

· it was necessary a big group of technical employees to operate the computer, but there were not enough trained people to hire.

Due to new technologies—hardware with RISC technology, UNIX operational system, RDBMS like Informix and Oracle, fast and cheap communication, etc—it is technically possible nowadays to distribute the information to several places.

Centralization, however had its advantages. There was a big centralized process to make decisions which big corporations liked. Now centralization has given way to allow several branches and departments more power in the decision making process. This occurred because decisions had to be made fast and shaped according to the

peculiarities of the regional market. Information had to be available in this decentralized way to guarantee its promptness when needed.

Because of these technical and managerial reasons, the proposal is to adopt a distributed database for managerial purposes. It will create and implement 18 regional databases, one in each DRT.

2.3 Confidentiality of the information.

A serious problem that results from the creation of a data warehouse is confidentiality. Who should be allowed to access the data base? How can access be controlled?

In order to solve this confidentiality problem we need:

- A good RDBMS that properly controls the allocation of rights to access the information.
- A Data Base Administrator that will be responsible for controlling the distribution of rights
- A well elaborated policy about rights of access.

At the beginning of this paper (item 1.2) it was said that the society has to have the right to access this Data Base. For confidentiality reasons, however, the taxpayer will only be allowed to access a subset of the complete Data Base. This subset will not have any specific information about individual companies, because the constitution protects fiscal secrecy.

Another important aspect to consider concerns the rights of each level of the hierarchy: the chiefs, inspectors and delegates must have complete right to access all detailed information about their jurisdiction and only aggregated information about other ones.

2.4 Executive Information System

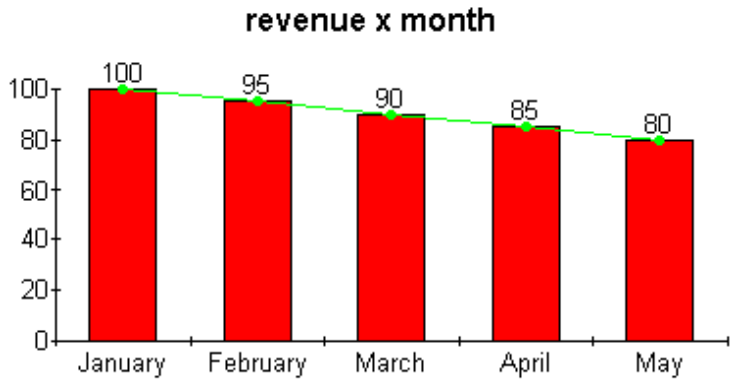
Computing is used in our daily lives for operational purposes; it is used in shops, banks, hotels, and other places. Operational operations are very common, but the most sophisticated usage of the computer is for managerial purposes. It provides the basis for correct decision making.

The data warehouse provides all information necessary for the EIS. After the building of the data warehouse the EIS is much easier. There is guarantee that it was used the same information base for every study made in every department of the organization.

The basic functions provided by the EIS are:

- trend analysis

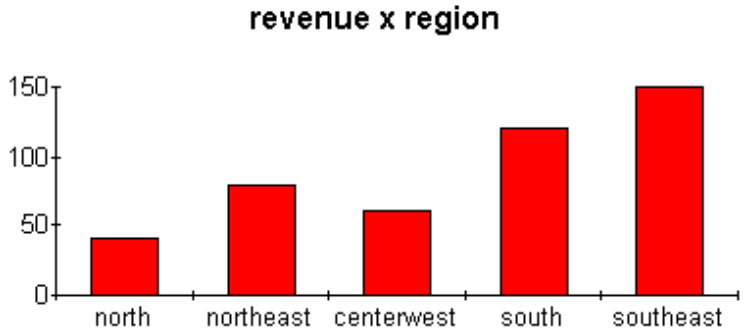
The EIS provides executive information about what the negative and positive trends are. The executive can then investigate the reasons for the trend.



slice and dice - multidimensional database

Another approach to analyze information is called "slice and dice." Using this concept the analyst is able to takes basic information from the data warehouse and group it one way to analyzes it, and then group it another way to reanalyze it. Slicing and dicing permits the executive to have different visions about the facts that are occurring.

For instance, he can analyze the facts according to month, quarters or years in the temporal dimension. After this, it is possible to analyze the facts according to regions of the country in the geographical dimension. The same graphic above would become as seen below:



To provide this feature two components are necessary: first, the data warehouse and then the Multidimensional Database.

The data warehouse is the foundation for the multidimensional DBMS. It feeds selected subsets of the detailed data into the multidimensional DBMS where it is summarized and otherwise aggregated.

The Multidimensional Database is a different way to see the information stored in a Database. We can imagine it as a cube, a named data cube with several dimensions. Each one of these dimensions is one aspect of the problem to be analyzed. For instance, we could imagine analyzing the tax revenue of the Sao Paulo Government according to time, region and taxpayer. These would be the three dimensions of the data cube.

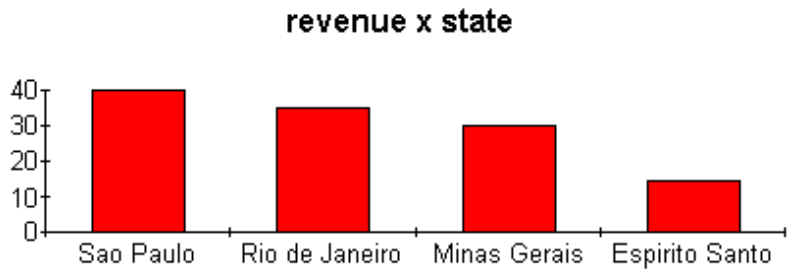
The Multidimensional Database provides high flexibility for accessing the data, according to the dimensions that were defined. In the example above, we could slice one period of time and analyze the information in the regions for this specific time. In so doing we are exploring the relationship between summary and detail data.

· drill-down analysis

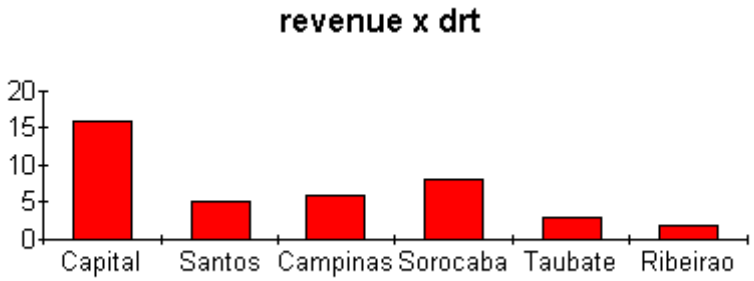
There is a useful saying that explains the necessity of a drill-down analysis: "first see the forest, and then see the trees." In order to understand the whole, we need first a global vision of the system. After, we need to see the components of the system to try to solve the problem that is occurring.

Drill down provides the capability to provide a global number, and then step by step break this number into a finer set of partial summaries according to the dimension that we want to study. The analyst by analyzing this sequence of values is able to understand what segment is causing the unexpected result.

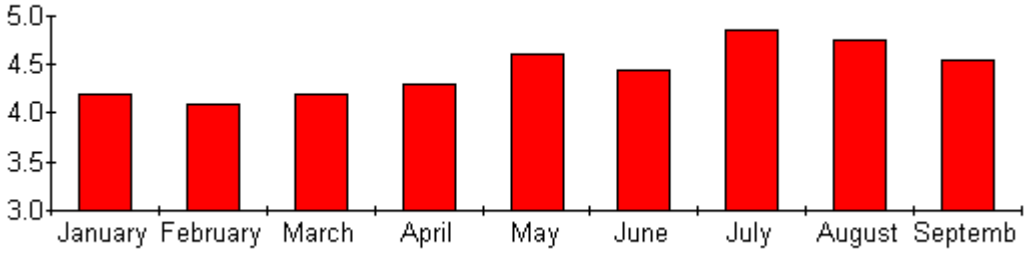
In the previous graphic, we may want to understand why the revenue of the southeast region is so high. To do this, we would request a detailed graphic about the southeast region. This graph would look something like this:



Again it may be necessary to request more details about this information. This request if made according to the DRTs would look something like this:



The same process could be applied for other dimensions of the analysis. In the temporal dimension, we could analyze according to year, quarter and month.



3. Deciding on the Data Base Management System and Software for analysis

3.1 Criteria for choosing the RDBMS

It was already said that the RDBMS is a key component in the whole process of Data Warehousing. The first step is to decide which one must be selected; therefore, we need a criteria to make the best choice of the RDBMS.

Based on the criteria suggested by H. Inmon in Building the Data Warehouse, there are four groups of aspects to be analyzed: Capability, open communication, administration, indexes and technical features .

3.1.1 Capability.

These are characteristics that permit software to manage a large volume and diversity of information stored in a data warehouse.

· manages great volumes of data

It is known that a data warehouse stores a large amount of data over a long period of time. Because of this fact the RDBMS has to be able to manage high volumes of data with high performance.

· manages multiple media

When we discuss data storage, we have to study the trade off between velocity and the cost of each media. In a data warehouse there is some information with a low level of access. This information can be stored in a media that is low cost and of low velocity. For information accessed more frequently, it must be stored in a high velocity media which is more expansive. The data warehouse administrator has to consider this aspect during the creation of the data warehouse.

The RDBMS must be able to permit all kinds of storage to be used simultaneously. It needs to make it so the user does not need to worry about where the information is coming from.

The media used for the memory (main, expanded or cache) can be DASD (disk or Direct Access Storage Device), magnetic tape, optical disk, and whatever emergent storage technique.

· manages data in parallel

A big change is happening in the hardware called parallel processing. Instead of a big machine with one processor, this process uses one machine with several small processors or several small machines working simultaneously. It is something like " small parts working together are as powerful as a big one".

This parallelism can be in terms of several processors or several physical devices to store data. This gives a boost in performance because the data may be read simultaneously on different devices. For instance, to read one million registers in a machine with one device will take tenfold the time as in a machine with 10 devices.

The RDBMS has to be able to store and process information in this type of parallel fashion.

· efficient loading of the data warehouse

It was already said that there are two environments: the transaction data base and the data warehouse. There is a great volume of data to be transferred between these two environments; therefore, the software must be very efficient in transferring this data. It is very useful if it can be done without paralyzing the processes in both environments. This process should be made in two phases: first the data, and then the indexes of the database in order to spread the workload evenly.

· quickly and completely restore data from a bulk medium

The data warehouse has to be on line 24 hours a day. When it is necessary to recover the information from a backup tape (due to whatever problem may have happened to the current data base), this has to be done as quickly as possible. Restoration from a tape, despite the low velocity of this device, must be made in a timely manner.

The software must provide an efficient feature to restore data from a tape.

3.1.2 Open communication

The data warehouse is exactly what the name indicates; it is a site of storage. It receives data from and passes data to a wide variety of sources. It must be able to provide the following features:

· Efficient and easy to use

The data warehouse has no use if it is not able to pass information to all kinds of software. This interface has to be easy to use, have few commands, and be simple to learn. This interface should be possible on a batch mode and on an on line mode.

· Rich language interface

Language is the tool that makes communication possible. The language of the data warehouse needs to be a robust language to easily insert, delete, update or access data. Nowadays the most common language is the SQL (Structured Query Language) which is a kind of Esperanto of the computer science. This language is used by almost all software.

These are some characteristics that are fundamental in this interface:

- can provide access to one record at a time

This is necessary when the user needs to know the value of only one specific record of the file.

- can provide access to one set at a time

This situation is more common, and happens when the user is not interested in one specific record but in the average or variance of a set of records.

- manages the indexes properly

Information has to be processed as fast as possible. Then the RDBMS has to decide how to access the information. It can be made going straight to the data or first reading a index. This decision has to be made by the software. The user is not interested in knowing how it is made. The plan to access the data has to be made by the software.

- provides an interface using SQL

Besides whatever language the software has, the SQL also has to be available.

3.1.3 Administration

All data has to be managed in order to be properly used. The main features to manage information are:

· Creation of a metadata control.

Metadata means data about the data. We need a lot of information about the data in order to store and use it properly. This metadata covers:

- The meaning of the data with a description of its contents.

- Origin of the data in the transaction process

- Formula used to calculate it

- Structure of tables

- Responsible for the updating the information

The data warehouse has to store this metadata accurately and keep it up to date.

· Manage rights to access the data (Security)

The data warehouse administrator controls the rights to access the data, including the rights to read or update the information. These rights may be given to persons individually to persons of one department or to persons of one post. The person who has one right may have a special right to give this right to another person which is called manager right (it is a kind of cascade right). These functions must be provided by the data warehouse in an easy and safe way.

3.1.4 Indexes

Due to the high volume of data, the technology used in the construction of indexes defines the efficiency of the data warehouse. These indexes delineate the best way to access the information. Here are some important aspects to be considered in this technology:

· Creation and monitoring of the indexes

The access of the data is unpredictable; nobody knows how it will be accessed by the user. The creation of indexes make access easier and faster. Fundamental decisions about how the indexes should be created, when they must be introduced, when they must be eliminated, and when they should be reorganized have to be made by the DWA. The efficient monitoring of these indexes has to be supported by the data warehouse, and changes in the index structure have to be easy and fast.

· Efficient index technology

Much technology exists about indexes, but it is not the objective of this paper to explain the index types. What must be stressed is that the data warehouse must efficiently use this technology. Here are some of the features:

- Partitioning the index in multiple disks

- having multileveled indexes

- storing the main part of the index in main memory, which uses the technology of very large memory.

- compacting the index entries when the order of the data allows it

-creating selective indexes and range indexes

- using bitmaps index (a type of index that indicates the existence of a condition for a record)

· Compound key

In a data warehouse it is necessary to register information over a relatively long period of time. In the transaction process it is registered only the information about one moment. An example of this would be finding the balance of an account. In the transaction process we need to know only the balance for the current moment. For the data warehouse we might need the balance for the last day of every month. Then we need a compound key which contains the month and number of the account.

A compound key is more common in a data warehouse, it has to support this feature.

· Index-only process

There are many questions that can be answered only by searching the indexes. It will be much more efficient if the query can be answered only by looking at the index. For instance, suppose we want to

know how many taxpayers there are in the city of Osasco. If an index is comprised of taxpayers by city, then it will be possible obtain the number of taxpayers only by reading this index. It will not be necessary to read the records of all taxpayers of Osasco.

3.1.5 Technical features

There are some technical features that are important for the efficiency of the system, which the data warehouse administrator (DWA) should know:

· Variable-length data

Due to the fact that there is a great variety of data in a data warehouse, and the information in a data warehouse is stable, (after the information is stored it is hardly changed) it is very useful to use variable-length data. In a transaction process it is not recommended to use this kind of data, because it can cause performance problems, but in a data warehouse there is no drawback to using this kind of data. The data warehouse has to support this data.

· Compaction of data

The ability to manage large amounts of data is important to the success of the data warehouse. Compact data can be stored in a minimal amount of space. In a transaction data base there is a problem with compaction of data because there is a huge overhead time for compaction and decompaction. In a data warehouse this is not a big problem, because the information is seldom updated which makes low overhead. The time to decompact is CPU time and not I/O time; I/O is more critical then CPU time.

3.1.6 The company that produce the software

Besides the characteristics of the product, it is very important to evaluate the company that produces the software. Some aspects must be considered in this decision:

· Support

It is important that the company have a good team of technicians to support the use of the software. This support must be available 24 hours a day.

· Line of products

To implement a data warehouse it is necessary to have several tools. It is appropriate that the software house have a comprehensive line of products to attend to every necessity. It should include DBMS, OLAP tools and CASE (Computer Aid System Engineering)

· Tradition

The software house has to have stability and longevity in the market. A data warehouse lasts for several years and it is very expensive to change the DBMS after the implementation of the data warehouse. Therefore, the software house must be trustworthy, and the DBMS and the software house must have a long life, as long as the data warehouse.

3.2 Criteria for choosing the tool for the analytical process (OLAP).

After the selection of the RDBMS it is necessary to select the software responsible for the processing of the analytical process. A criteria should be established to conduct the selection of the most appropriate software.

The important criteria can be divided into three big groups: essential, necessary and technical.

3.2.1 Essential

· This software has to provide a multidimensional view of the data, according to the several dimensions defined in the model of data (star schema).

· In the earlier chapters it was seen that our data warehouse is based on a Relational Data Base Management System (RDBMS). Therefore, the OLAP tool has to be a software that accesses data in a RDBMS.

· The communication between the OLAP tool and the RDBMS may be made by using a process named Open Data Base Communication System (ODBC) or by using a native communication between them.

3.2.2 Necessary

Besides these fundamental characteristics, there are other aspects that must be considered in the decision process.

· Partitioned.

It has to work in a partitioned fashion. This means it has to divide the process into two parts, one in the central host and another in the client computer. Partitioning aims to reduce the traffic of data in the network and use of the central computer. In order to reduce the traffic of data in the network, the software has to process all bulk information on the server computer, and to reduce the use of the server computer it must process and format the summarized information on the client computer

· Portability.

This software must be able to access the data warehouse, regardless of the environment of the data warehouse; i.e., the software must be able to access a UNIX or Windows NT environment. These are two kinds of operational system that can be simultaneously present in the final integrated solution adopted by the information system.

· Integrated.

This solution has to provide an integrated analytical environment. The access to the data warehouse, the creation of graphics, and the creation of spreadsheets are made using software or modules of the same softhouse.

· Extensibility.

It has to be possible to add the rules and functions of the businesses of the organization in the multidimensional analysis of data. The software must make it possible to model the analysis according to the process used in the company.

· Sample

It is known that the data warehouse has a bulky quantity of information; therefore, processing this information usually takes a long time. Analysis is an interactive process. Several attempts may be necessary before the final result is reached. In order to solve the problem of accessing a large quantity of information several times, which would take a really long time, a sample should be used. In doing so, it is possible to simulate several possibilities in a shorter period of time. The final result may then be processed using the whole database.

Based on this explanation, we can see that the sample process is a very important feature that must be available in the software.

3.2.2 Technical

· Possibility to include small routines named "plug-ins" using the language C++. These routines are very efficient and make possible the creation of functions that define the rules of the business.

· Resourcefulness to execute these functions: "drill up", drill down" and "drill across" in a spreadsheet. "Drill up" means the ability to summarize the data one level above the present level; "drill down" means to detail the data one level below the present level, and "drill across" means to analyze the data in a segment beside the present level.

· Permits the user during the analysis to create dynamic groups, using any or several of the existent dimensions (variables). For example, the user can create a group encompassing every person between 18 and 31 years old that has a yearly income between $25.000 and $40000.

· Permits the creation of graphics, queries and spreadsheets in the form of tables, making it possible to switch quickly between formats.

· Permits a query to be completely executed on remote equipment (Server), and delivers to the client only the final result.

· Permits the creation, updating and utilization of pre-calculated data (aggregated) in order to enhance the performance of the system. This procedure must be transparent to the user.

· Permits an open interface between the software and the EIS available in the market. This interface must be possible with MS-Excel as well as other electronic spreadsheets.

· Permits the administration of metadata, using a graphic tool. This administration must complement the metadata used by the RDBMS.

· Permits the creation of incremental aggregates. This means that all new or updated data will automatically update the corresponding statistics.

· Resourcefulness in analyzing the RDBMS for establishment of a strategy to create aggregated values.

· Capacity to create samples according to the parameters defined by the users.

· Implementation of security rules to define and execute queries. This is necessary in order to restrict who has the right of creating and executing queries.

· Contains the feature to statistically analyze the most used queries and distribution of data. This will aid the database administrator (DBA) in the process of creating aggregates. In this way, the DBA will optimize the access time and efficiently use disc space.

· Ability to screen and modify the angle to visualize the data ("slice and dice").

4. Modeling a first version of an Analytical Data Base

This chapter describes the first version of a model of the data warehouse about collection of taxes. This model uses the proper methodology of building a data warehouse; but, it is a small and simplified representation of the complete solution that is expected to be implemented.

Due the fact that the data warehouse is constructed in an iterative fashion and that the requirements for the data can not be known a priori, the model presented here is only a first and rough version.

This model will be the framework for the correct construction of a data warehouse, and for the efficient use of OLAP.

4.1 Criteria to build a data warehouse (measure, dimensions, granularity and star schema)

Before showing the proposed model, it is necessary to explain the steps necessary to build it. There are six steps that must be followed:

· Objective.

The first step is to define the objective. This objective represents the aim and the comprehensiveness of the model.

· Queries to be answered.

Second, compile a list of the queries that must be answered by the model. This list must consider every level of the organization it will be expected to serve.

· Dimensions to be considered.

Third, every analysis has a central entity where the queries are submitted. This central entity is called a "fact table." This fact table will be heavily populated and will be related to all the surrounding tables. The surrounding entities are called "dimension tables." These are not heavily populated and represent the levels (dimensions) where the information will be analyzed.

· Measures.

After defining the fact table, the fourth step involves delineating what variables must be placed in this table. These variables are called measures and consist of the information necessary to answer every question of the query list.

Each element of the fact table has a group of data necessary to satisfy the query list.

· Granularity.

"Granularity refers to the level of detail or summarization held in the units of data in the data warehouse. The more detail there is, the lower the level of granularity. The less detail there is, the higher the level of granularity." (H. Inmon)

The data in the fact table must be summarized as much as possible. In doing so the volume of registers is reduced and therefore the performance is enhanced. This data cannot be summarized if any question of the query list becomes impossible to answer. In other words, the fifth step involves summarizing the fact table as much as possible with the query list acting as the limit.

· Star scheme.

After defining the fact and dimension tables, we must represent graphically the relationship between these tables. This graphic is called the star scheme, and is the sixth step.

4.2 Information about tax collecting (Data Mart)

The data mart is a subset of the data warehouse. The source of all departmental data is the data warehouse, the departmental level is called data mart.

In this chapter the steps described above are applied to constructing a data mart about tax collecting.

This chapter is based on a project jointly made with Informix.

4.2.1 Objective

In the Finance Secretariat of Sao Paulo, it was decided that one of the most urgent areas information to be analyzed was tax collecting of the ICMS. Therefore, it was decided to implement a data mart about ICMS levying. This was done in the following manner.

4.2.2 Queries to be answered.

Before we study the queries to be answered, it is necessary to explain that all information used in this data mart comes from the documents GIA and GARE. These documents are explained in 1.2.3.

It is also necessary to explain in more detail how the calculation of the due ICMS is made. The due tax is equal the debit tax (the tax referred to the sales) less the credit tax (the tax referred to the purchases).

The questions to be answered were divided into four categories, according to the hierarchical level of the manager (Secretary, Coordinator, Director and Delegate).

4.2.2.1 Secretary

· Monthly collecting

This is an analysis of the revenue collected monthly. For instance we can know: monthly variation of the collecting, aggregated values in the period, comparison between two periods in different years. This value comes from the document GARE, in the field total value (a_collecting).

· Payment default index

This is the percentage between the due value informed in the GIA (field a_saldo_devedor) and the paid value informed in the GARE (field a_valor_icms).

· Collecting by economic sector

This is an analysis of the collecting according to the CAE (Codigo de atividade economica - Economic activity code). The 5 major economic activities in the CAE are: industry, wholesale, retail, services and miscellaneous.

4.2.2.2 Coordinator.

· List of the 100 mayor taxpayers according to the taxes paid. Field total value (a_arrecadacao) on the form GARE.

· Collecting by economic sector.

This is the same analysis as the secretary, but with more details. Every sector will be analyzed in a lower level of aggregation; i.e., each sector will be analyzed according to the kind of business done.

· Payment default index by economic activity.

4.2.2.3 Director

· Collecting by DRT, monthly variation and participation of each DRT in the total value. This information is in the field total value of GARE (a_arrecadacao).

· Default in delivering GIA by DRT.

This is the number of taxpayers that did not deliver the GIA for the current month, and this value has to be analyzed on a monthly basis.

· Payment default index by DRT.

· Number of taxpayers with persistent surplus balance in the GIA. Persistent means more than three months straight with surplus. Surplus balance means that the taxpayer is buying more than selling, i.e., he does not have added value. This situation is a sign that he is evading taxes.

· Collecting aggregated by company. A company has several branches, all of which have the same basic CGC (Cadastro Geral de Contribuintes - CGC is the Taxpayer number). It is important to analyze the company on an aggregated basis.

4.2.2.3 Delegate

· Collecting by IF, monthly variation and participation of each IF in the total of the DRT.

· Number of taxpayers and quantity of collecting according to :

· Economic activity

· Class - the taxpayers according to ranking in the total. There are only 4 classes: A, B, C, and D.

· Regime - the responsibilities the company has to the government, according to its business profile. For instance, micro companies do not have to deliver the GIA monthly, only yearly.

· Tax/income index according to economic activity.

Indicates what the level of the tax based on the revenue of the company is, and is useful when comparing the index of the company with the index of the economic sector where the company works.

· Taxpayers with debit tax divided by due tax less than 5%.

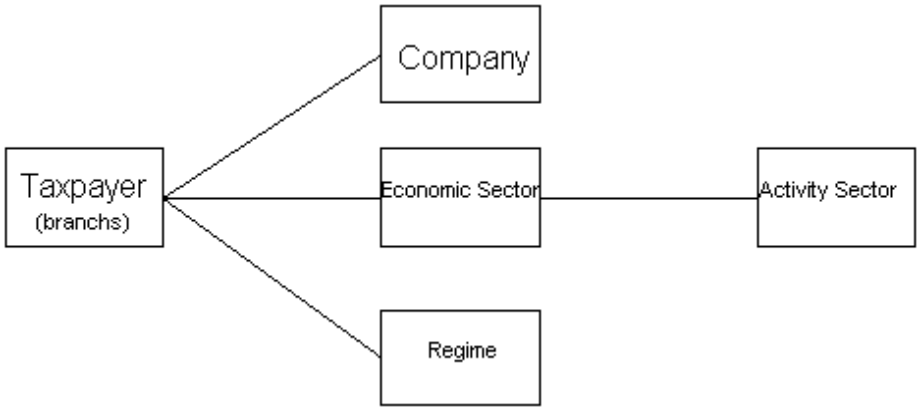
Important to measure because it represents the level of added value. If it is too small, defraudation may be occurring.

4.2.3 Dimensions to be considered.

Based on the queries, it was decided to create three dimensions.

· Taxpayers: Information about every company that pays tax.

This dimension has this hierarchy.



A taxpayer is any branch of the company , or the entire company in cases where the company does not have a branch. The taxpayer is identified by the whole CGC (12 digits).

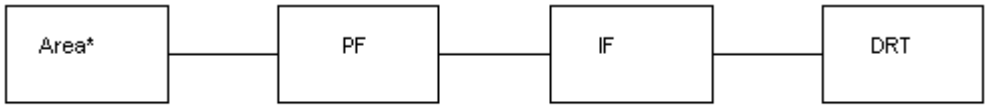
The taxpayers may be aggregated by the headquarters which is identified by the 8 initial digits of the CGC. They can also be aggregated by the economic sector, and the economic sectors are aggregated by the activity sector. Finally, the taxpayers can be aggregated by regime and by class.

These are the fields of the dimension taxpayer:

Field	Number of characters
c_ie	12
c_cgc	8
c_name_company	45
c_cae	2
c_name_cae	60
c_sector	1
c_name_sector	15
c_regime	3
c_name_regime	15
c_class	3
c_name_class	15
c_name_branch	45

Geography: Represents the geographical distribution of the payments.

This dimension has this hierarchy.



*(city or part of a city)

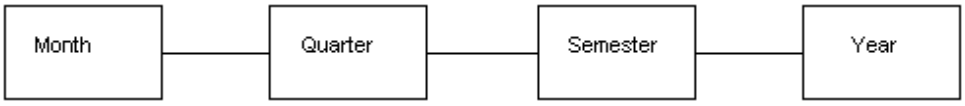
An area is commonly a city. Sao Paulo city is the only exception; due to the large size of the city, it was divided into 3 areas. Each area is aggregated into PF which is then aggregated into IF, and finally aggregated into DRT.

These are the fields of the dimension geography:

Field	Number of characters
g_area	6
g_name_area	20
g_pf	4
g_name_pf	15
g_if	4
g_name_if	15
g_drt	4
g_name_drt	15

Time: Represents the way the data is analyzed in relation with time.

This dimension has this hierarchy.



These are the fields of the dimension time:

Field	Number of characters
t_time	6
t_code_month	6
t_name_month	10
t_code_quarter	2
t_name_quarter	16
t_code_semester	1
t_name_semester	16
t_year	4

4.2.4 Measures to be used.

These measures represent the data to be recorded in the fact table. In this study, the fact table is the collecting table.

Based on the query list, the following measures were created.

Fact Table - Collecting

Measures	Source (field)	Name
Debit of tax	55 from GIA	a_debit_tax
Credit of tax	62 from GIA	a_credit_tax
debit balance	65 of GIA (if negative)	a_debit_balance
credit balance	65 of GIA (if positive)	a_credit_balance
income	74 from GIA	a_invoice
collected tax	14 from GARE	a_collected
ICMS paid	9 from GARE	a_icms
default delivering GIA	1-delivered, 0-not delivered	a_gia

4.2.5 Granularity to be used.

The granularity to be used will be an entry for each month for each taxpayer. In the case of a taxpayer that has two payments (GARE) in the same month, both will be totaled and will be only one register of the fact table collecting.

4.2.6 Model to be used.

Based on the dimensions and the measures identified in this study, this is the proposed star schema.

5. Advanced tools for an analytical process (data mining)

5.1 What is data mining?

The earlier chapters explained how a data warehouse provides the organization with memory, but learning requires more than simply gathering data. After creating the data warehouse, the data must be analyzed, understood, and turned into actionable information. This is where data mining enters.

Data mining is a set of tools that adds intelligence to the data warehouse. Using these techniques, it is possible to exploit the vast mountains of data generated by interactions with the taxpayers. Also we are able to make prospects on this information in order to know the taxpayers better. Using data mining, we can ask who is likely to pay taxes correctly and who is likely to evade?

The goal of data mining is to allow a corporation to improve efficiency and efficacy, as well as support operations through better understanding of its customers. These techniques can be equally applicable in fields ranging from law enforcement to radio astronomy.

Some tasks well suited to data mining are classification, estimation, prediction, affinity grouping, clustering and description.

This type of analysis can be done in two ways. The top-down manner is called hypothesis testing. In hypothesis testing, a database recording past behavior is used to verify or disprove preconceived notions, ideas and hunches concerning relations in the data.

Other tasks are best approached in the bottom-up manner called knowledge discovery. In knowledge discovery, no prior assumptions are made. The data is allowed to speak for itself. Knowledge discovery comes in two forms, directed and undirected. Directed knowledge discovery attempts to explain or categorize particular data fields, such as income or response. Undirected knowledge discovery attempts to find patterns or similarities among groups of records without the use of a particular target field or collection of predefined classes.

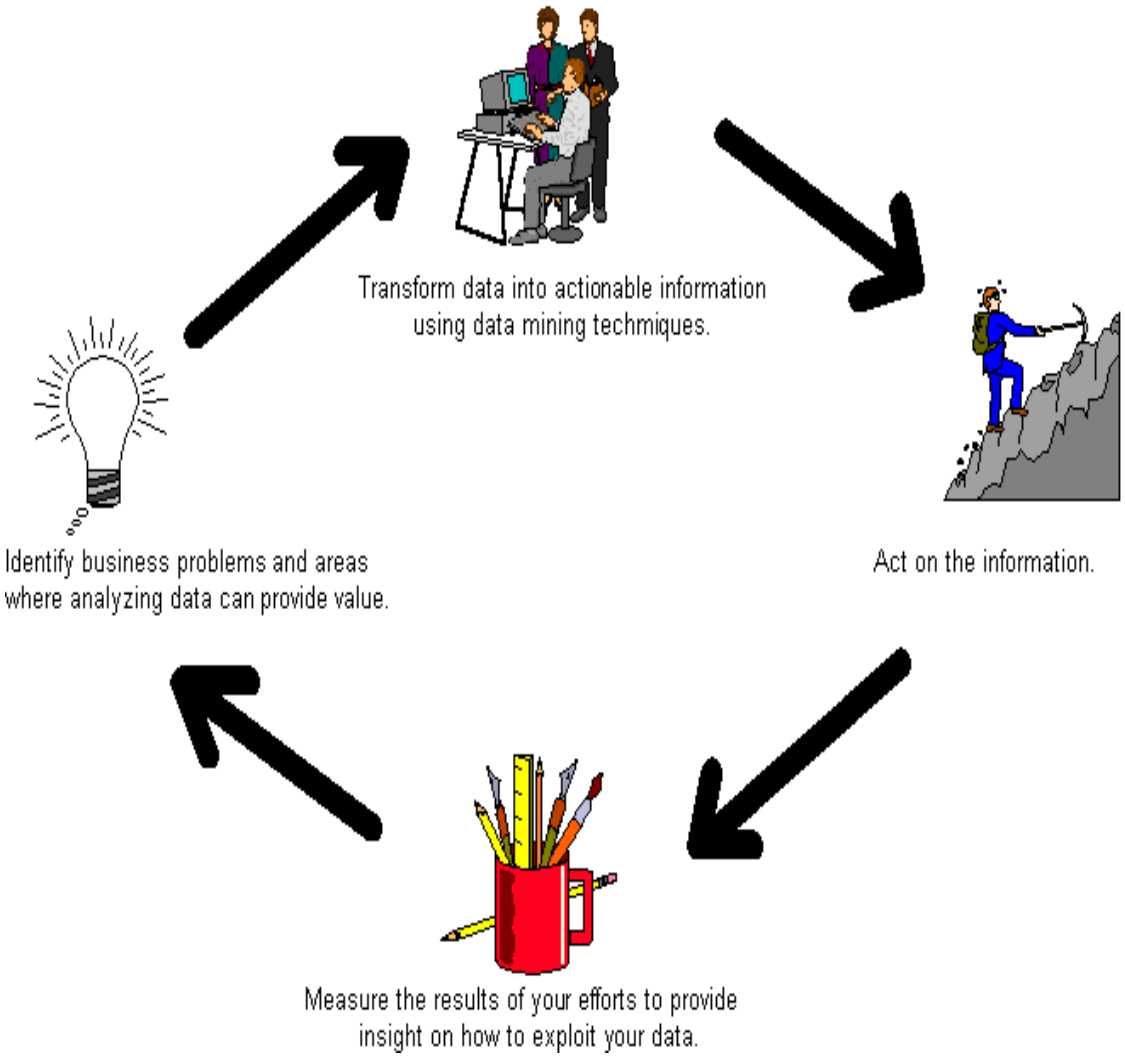
5.2 The virtuous cycle of data mining

data mining works to find the interesting patterns lurking in all those billions of bytes of internal and external data. Merely finding the patterns, however, is not enough. We must be able to respond to the patterns, to act on them. Then, we need to be able to turn the data into information, the information into action, and the action into value. This is the virtuous cycle of data mining. It focuses on action based on discovery, not on the discovery mechanism itself. Success in using data will transform an organization from reactive to proactive.

It seems that data mining will solve every problem of the organization, but that is not an easy task. Part of the solution is based on data mining, but much is also based on the organization’s experience and knowledge about the business. The organization has to be able to identify opportunities and turn the results from the data mining into action.

The virtuous cycle is divided into four stages:

- Identify the business problem.
- Use data mining techniques to transform the data into actionable information.
- Act on the information
- Measure the results



5.2.1 Identify the business problem/opportunity.

Every problem is a real opportunity to improve the performance of our activities. This is the reason that both terms are together.

In this step, the purpose is to identify the areas where data can provide value. Identifying business opportunities or problems occurs throughout the organization, wherever increased information enables people to better perform their jobs.

In our case of tax collecting, we already rely on some amount of data analysis while during our operations. It can be treated as the Identify stage for the virtuous cycle. Here are some examples:

- Planning the efforts of the tax auditors
- Targeting the action of the auditing
- Understanding the taxpayers' actions
- Analyzing the levying based on regional and economic sector

5.2.2 Data mining

A data warehouse must be created before starting the process of data mining. The data warehouse organizes in one area the data that comes from multiple systems in many formats. Creation of the data warehouse must occur before data mining can occur.

The way how to use the data mining techniques will be depicted in the following topics in this paper.

5.2.3 Take action.

This is where the results from data mining are acted upon, and results are fed into the measurement stage. It is easy to understand that the action has to be based on the information to better understand the problem and be well prepared to face it. But, (there is always a but) the question is how to incorporate information into the business processes, so the actions are an integral part of the virtuous cycle.

It is necessary when implementing new strategies and policies, to be sure to collect data needed to understand the effects of new policies. For successful data mining, the business processes must provide the data feedback needed for the virtuous cycle.

We need to change the mindset of operational groups. Successful results, when measured and communicated, will encourage other groups to start incorporating data-driven information into their processes.

5.2.4 Measure the results.

Measurement provides the feedback for continuously improving results. It is the same for organizations or at the individual level. We improve our efforts by comparing and learning.

Usually there are reports measuring the results of the operation. The problems are:

· The information included in the report

· The timeliness of the information

· Availability of the information

In general these reports are produced a long time after the event, and are not analyzed carefully by the managers. It’s essential to analyze the results of the action before the action is over, in order to do this the fast production of information reports is crucial.

Data mining is an interactive process. Every data mining effort is like a small business case. By comparing our expectations to the actual results, we can often recognize promising opportunities to exploit on the next round of the virtuous cycle. Usually, we are too busy thinking about the next problem instead of measuring the success of the current efforts.

In order to correctly measure the effectiveness of the data mining efforts, we depend on information provided in the previous stages. It is crucial to ask the right question early, in order to collect the right information for measurement.

5.3 The tasks data mining can do.

How to loose 10 pounds in one week is not the kind of task that data mining will solve. It is not a panacea for every kind of problem, which is the unrealistic expectation of several managers. Most business problems, however, can be solved using one of these six tasks: Classification, Estimation, Prediction, Affinity grouping, Clustering and Description. These tasks can be solved using data mining techniques. Let’s briefly explain these tasks based on the definitions given by Michael J. A. Berry and Gordon Linoff in Data Mining Techniques:

· Classification.

Classification consists of examining the features of a newly presented object and assigning it to one of a predefined set of classes. The classification task is characterized by a well-defined definition of the classes, and a training set consisting of pre-classified examples. The task is to build a model of some kind that can be applied to unclassified data in order to classify it.

· Estimation.

Estimation deals with continuously valued outcomes. Given some input data, we use estimation to come up with a value for some unknown continuous variables such as tax, income, height, etc.

· Prediction.

Prediction is when the records are classified according to some predicted future behavior or estimated future value. In a prediction task, the only way to check the accuracy of the prediction is to wait and see.

· Affinity grouping .

The task of affinity grouping is to determine which things go together, The prototypical example is determining what things go together in a shopping cart at the supermarket.

· Clustering.

Clustering is the task of segmenting a heterogeneous population into a number of more homogeneous subgroups or clusters. What distinguishes clustering from classification is that clustering does not rely on predefined classes. In clustering, there are no predefined classes and no examples. The records are grouped together on the basis of self-similarity.

· Description

Sometimes the purpose of data mining is simply to describe what is going on in a complicated database in a way that increases our understanding of the people, products, or processes that produced the data in the first place. A good enough description of a behavior will often suggest an explanation for it as well.

5.4 Data mine techniques.

data mining is part of the virtuous cycle. After the careful analysis of the problem we can use data mining techniques. These techniques are briefly explain below and are based on the definitions given by Michael J. A. Berry and Gordon Linoff:

· Market basket analysis

It is a form of clustering used for finding groups of items that tend to occur together in a transaction. The models that it builds give the likelihood of different products being purchased together and can be expressed as rules.

· Memory-Based reasoning (MBR).

MBR is a directed data mining technique that uses known instances as a model to make predictions about unknown instances. MBR looks for the nearest neighbors in the known instances and combines their values to assign classification or prediction values.

· Cluster Detection

This is the building of models that find data records that are similar to each other. These clumps of self-similarity are called clusters. In tax auditing, this technique could study what groups of taxpayers have the same pattern of paying.

· Link analysis.

Link analysis follows relationships between records to develop models based on patterns in the relationships. In tax auditing it can analyze who the common supplier is of each tax payer. It can analyze if tax evasion is occurring in these operations.

· Decision tress and rule induction.

In this model the records of a training set are divided into disjoint subsets, each of which is described by a simple rule on one or more fields. The new records are classified according to these simple rules.

· Artificial neural networks.

These are models of simple neural interconnections in brain adapted for use on digital computers. In their most common incarnation, they learn from a training set, generalizing patterns inside it for classification and prediction.

· Genetic algorithms (GA).

GA applies the mechanics of genetics and natural selection to a search for finding the optimal sets of parameters that describe a predictive function. GA algorithms use the selection, crossover, and mutation operators to evolve successive generations of solutions. As the generations evolve, only the most predictive survive, until the functions converge on an optimal solution.

Here is a chart that explains which technique can be used according to the task to be done:

Technique	Classification	Estimation	Prediction	Affinity Grouping	Clustering	Description
Standard Statistics	X	X	X	X	X	X
Market Basket Analysis			X	X	X	X
Memory-Based Reasoning	X		X	X	X	
Genetic Algorithms	X		X			
Cluster detection					X	
Link analysis	X		X	X		
Decision Trees	X		X		X	X
Neural Networks	X	X	X		X	

6.0 Conclusion.

In this paper it was showed how important is the study of the huge amount of data that the Information Technology has made available for the use of the managers. It is important to improve corporate efficiency through data driven decisions.

It was shown that the first step in organizing the information is a data warehouse. The second step is to introduce the new way of making decisions as shown in the virtuous cycle of data mining. In doing so, we are able to extract better results using information technology, and therefore improve the efficacy and efficiency of organizations.

In the Finance Secretariat of Sao Paulo we started this process. But, we are only in the beginning of this process, there are many things to be done. We have to implement several data warehouses and after this start the process of data mining. It will be a long process, but it is worth the budget and the sweat that will be spent on this project. Implementation of this proposal will make it possible for us to give better service to the community.

Understanding data is not a new aim in the history. What makes the latter twentieth century different is the information technology which makes possible the analysis of huge amounts of data.

References

Data mining Techniques

Michael J. A. Berry and Gordon Linofff

Building the data warehouse

W. H. Inmon

Data warehouse - First model for Tax collecting

Informix and Finance Secretariat of Sao Paulo

Advanced Topics in Information Systems - Handouts

Prof. B. Narahari