**Thematic Week:** Water Economics and Financing

**Thematic Axis:** Financial Solutions for Emerging Countries

**Title**: Business intelligence for Water: The limits on information

**Authors**: José Gordon Rapoport

**Abstract:**

Private sector institutions like Hospitals, Banks or Manufacturing companies have since long benefited from Business Intelligence tools. This tools comprise analysis, reporting, scorecarding and even datamining components which make possible a better understanding of the business by allowing the processing or digestion of several thousands of individual pieces of data. The goal of this presentation is to introduce this type of technologies into the Water Resources management institutions. A brief explanation of concepts like multidimensional OLAP analysis, Datamart, and Data mining algorithms is given.

**Keywords:** Business Intelligence; KPI; Scorecard; Data Mining; Datamart; Decision Support Systems

**Introduction**

Each day any institution gathers or processes more data about its activities, detailed data that the human brain is unable to comprehend in its raw form unless it is grouped and classified in perspectives or dimensions.

The Business Intelligence concept was born as an answer to a common concern: How are we going to take advantage of the huge amounts of data that grows from day to day? How is it going to be analyzed? How can we use it efficiently?, How can technology help us in order to generate intelligence?, How can we get fast answers to our questions and how can data help in the decision making process?, How can the level of uncertainty be minimized?

Besides that, the corporate world has seen a steady shrinking in the headcount of the Office of Finance, so more had to be done with less human resources.

Transactional systems, and specially ERP's (Enterprise Resource Planning systems) had been of great help for the Office of Finance, they have simplify the work of the accountants and have made possible leaving behind the world of multiple systems with duplicated and even triplicated data, and arriving to the integrated-do-it-almost-all system. In these integral systems a transaction has an effect over different areas of the institution, for example an invoice may affect inventory, acounts receivable, taxes, etc.

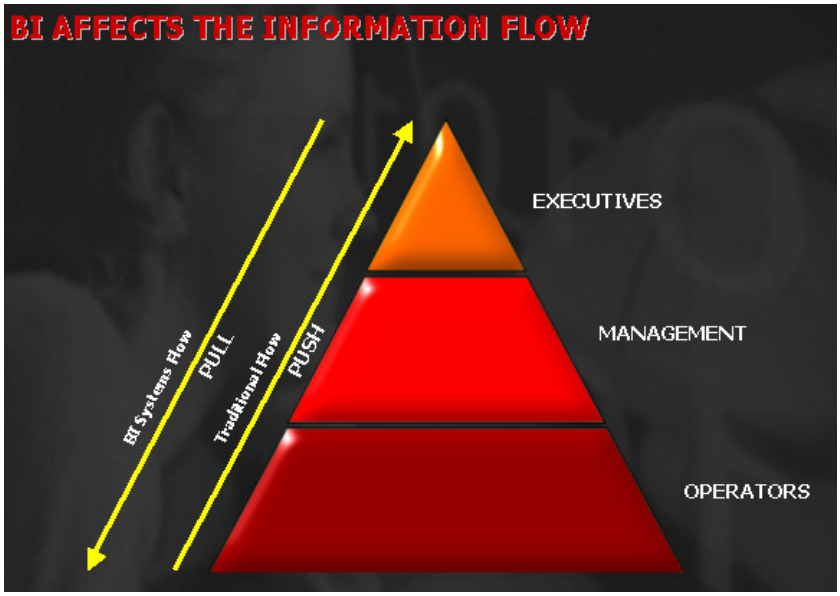
These systems are oriented toward being very efficient when registering a transaction, but rather bad at presenting aggregated information, since they are not designed to store it, but each time that a subtotal is required it has to be calculated at the expense of affecting the whole system performance.

And here is were the key to the success of the Business Intelligence tools lays and it can be explained in terms of two main factors: the heavy overload that an aggregation query has on the transactional systems; and the huge amount of data that has to be analyzed according to multiple perspectives in order to make a decision. Each one of those perspectives is a dimension. The when, where, in which market, are the kind of perspectives that we are talking about.

So the perfect combination is: ERP for transaction registering and Business Intelligence –from now on BI- systems to present or use the information generated.

Generally speaking, Business Intelligence is a set of applications and technologies used to gather data, enable access and analyze data about the transactions or operations of an institution in order to supply information that will support better decision making.

In the traditional flow – without BI systems- the Operative people generate reports that are pushed up to the Management and the High Direction. BI shifts the paradigm and enables the High Direction to look directly at the information and make whatever analysis they like without intermediaries, because BI tools pre-sumarize the data so the users can make unplanned queries rather quickly.




Business Intelligence systems provide:

- a multidimensional approach
- transform data into information
- a guided tour into the bunch of data
- an easier correlation of facts
- answers to business related questions

Although there are a lot of variants of BI components that cover a wide range of applications, this can be classified in two main classes:

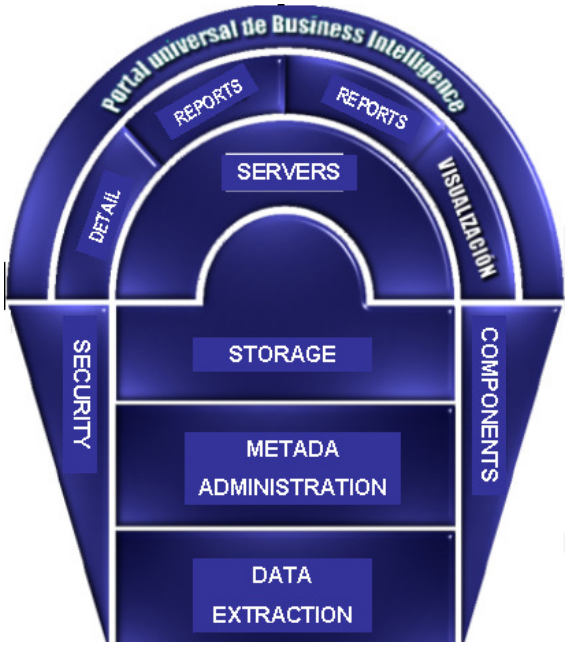
- Query Reporting and Analysis (QRA)
  - Ad-Hoc Query
  - OLAP Analysis – or Multidimensional Analysis
  - Reporting
  - Scorecarding
- Advanced Analytics.
  - Datamining
  - Statistics

The application of this type of systems is widespread in the Corporate World, there can be found some experiences in Government Agencies, and there are almost no experiences in Water Management agencies. An internet research has yield some hits on BI deployments in the Denver Water Authority, Las Vegas Valley Water District, Soth East Water and Sidney Water in Australia.

Many Water related agencies have turned their eyes into the use of Geographical Information Systems (GIS), which are natural complements of Business Intelligence Systems. They are good at displaying information in a map about a certain attribute of specific points or areas, but they lack the ability to make comparissons in time or against targets and the whole concept of slice and dice, drill down, drill thru or dimension swapping is missing in them.

In the following pages we are going to explore the BI world, some of its most known tools and some basic concepts, but the main goal is to have in mind how this can be of use in the water related sector.

**Components of a Business Intelligence solution**




A typical BI application takes data from the transactional or operational systems, an even from unsystematized sources as spreadsheets or text files. In order for the information presented in the BI systems to be reliable, automated processes must be developed using ETL tools – short for Extraction, Transforming and Loading. This tools clean the data and perform heavy load processing.

The transactional systems used by the institutions are designed having in mind efficiency at the transaction level, but they are rather clumsy at sumarizing since they don't store aggregations, so they have to sweep all the database in order to calculated them, thus affecting the system performance and slowing it down, even to the point of breakdown when several of this queries are requested by multiple users. That is why the data has to be taken from the ERP into a specialized repository – the datamart- through the ETL tools.

The datamart is also the natural gathering point for data coming from diversity of sources: Excel files, HR systems, Fixed Assets, ERP system and others. The data that lands here is assumed to be homogeneus, clean and consistent.

The different aspects, perspectives or dimensions by which data is analyzed have to be consistent along the datamart. For example, the way data is regionalized should be the same if we are talking about weather stations or if water quality data is displayed.

A sound datamart should have Metadata Management, which refers to the collection of what each field represents, from where does it feed, filters that have to be applied and any other relevant information about each piece of data. Literally metadata means "information about the data".

Security is a basic component since the information has to be viewed only by those authorized to see it.

The datamart is really a set of related tables that seat together in a database. This tables are organized and prepared to facilitate the construction of reports or of OLAP cubes. This preparation involves the definition of star or snowflake schemas. The tables that contain the data are refered as "Fact Tables", where many data fields can be found, plus fields that will be used to build the hierarchies or dimensions, that is, the different business perspectives. The star or snowflake schemas use additional tables related to the Fact Table in order to build the complete hierarchy.

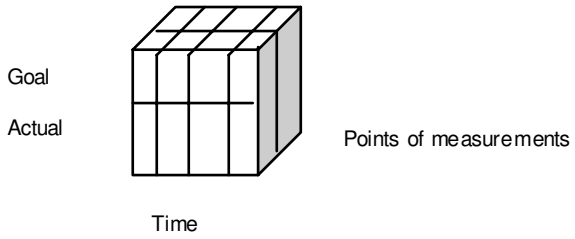
On top of everything seats a Web portal thru which the users have access to the different reports. There are different kinds of reporting strategies: Ad Hoc query; predefined reports; or OLAP cubes.

The ad-hoc query builders allow the end user to look directly at the information stored in the datamart, thus enabling him to get a list of any available fields. Also, he could have access either to predefined reports that run with up-to-date data or to OLAP cubes where he can play with the dimensions.

**Multidimensional Analysis**

Multidimensional Analysis is also known as OLAP – (On-Line Analytic Processing). This concept engulfs a set of technologies that enable the user to analyze data in a multidimensional way, thus enabling him to make any type of comparissons of preprocessed aggregations. OLAP Analysis is strictly tied to the word Cube or Hypercube, since each one of the aspects of the business is represented by a side of the cube or dimension.

For example, assume that Water Quality Index data has been integrated into a Cube. One dimension would be the "where" that is the point where the measurements were made. The time dimension would be the "when", that is the date of measurement. In the same set we would like to have the "version" dimension in which we could compare the actual value against the desired or allowed value and a variance in percentage. Another dimension would be the measurements dimension, that comprises all the related parameters measured for that same point, in the same date, for actuals and for target values.




In spite of its name, analysis aren't done really on-line but on previously processed data. OLAP systems are used in the Corporate world to make easier for managers and analysts to monitor the business performance.

Cubes or hypercubes analysis give much help in:
–Abnormal circumstances detection
–Problem understanding.

OLAP Components

In order to build and Olap system there are some requisites, like:

•Clean and reliable datasources. There is no point in having isolated sources, there should be a single datasource from where all the analysis come, this are normally provided by the ERP or transactional systems.

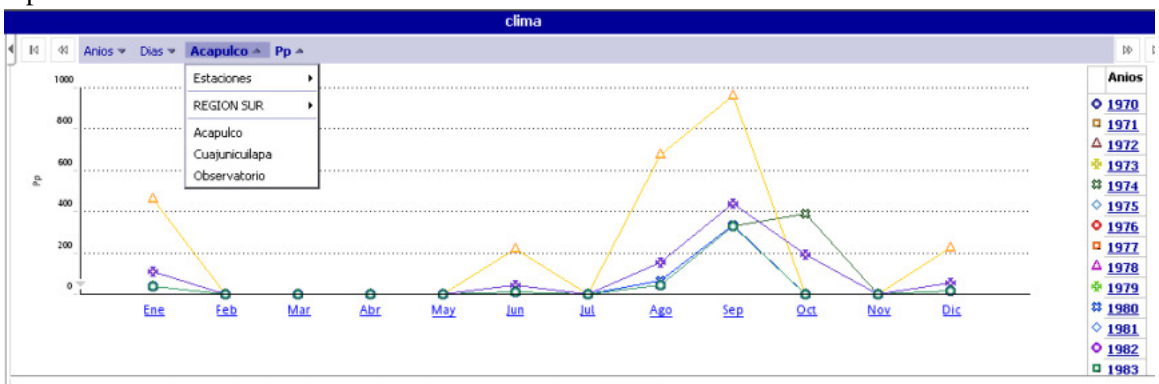
•Tools for the automation of data extraction (ETL). The final user of OLAP assumes that the the displayed data is an updated and reliable one. Extraction automation is an essential part of this process and is normally based in specialized tools.

•Datamart, Datawarehouse or data repository. The extracted data is placed in a special repository who is only altered or written by the ETL tools.

•Mulditimensional cube creation. There are many commercial tools that aggregate the data into subtotals and actually create the cube, among these we find:

–Analysis Services (Microsoft)
–Transformer (Cognos)

•Cube publishing. This is normally done through a Web server, usually Microsoft's IIS or Apache




| Pp como valores | Ene | Feb | Mar | Abr | May | Jun | Jul | Ago | Sep | Oct | Nov | Dic | Dias |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1970 | 38.87 | 0.00 | 0.00 | 0.00 | 0.00 | 9.91 | 0.00 | 43.69 | 337.06 | 0.00 | 0.00 | 16.00 | 445.53 |
| 1971 | 38.87 | 0.00 | 0.00 | 0.00 | 0.00 | 9.91 | 0.00 | 43.69 | 337.06 | 0.00 | 0.00 | 16.00 | 445.53 |
| 1972 | 38.87 | 0.00 | 0.00 | 0.00 | 0.00 | 9.91 | 0.00 | 43.69 | 337.06 | 0.00 | 0.00 | 16.00 | 445.53 |
| 1973 | 38.87 | 0.00 | 0.00 | 0.00 | 0.00 | 9.91 | 0.00 | 43.69 | 337.06 | 0.00 | 0.00 | 16.00 | 445.53 |
| 1974 | 38.87 | 0.00 | 0.00 | 0.00 | 0.00 | 9.91 | 0.00 | 43.69 | 337.06 | 400.00 | 0.00 | 16.00 | 845.53 |
| 1975 | 38.87 | 0.00 | 0.00 | 0.00 | 0.00 | 9.91 | 0.00 | 43.69 | 337.06 | 0.00 | 0.00 | 16.00 | 445.53 |
| 1976 | 38.87 | 0.00 | 0.00 | 0.00 | 0.00 | 9.91 | 0.00 | 43.69 | 337.06 | 0.00 | 0.00 | 16.00 | 445.53 |
| 1977 | 38.87 | 0.00 | 0.00 | 0.00 | 0.00 | 9.91 | 0.00 | 43.69 | 337.06 | 0.00 | 0.00 | 16.00 | 445.53 |
| 1978 | 38.87 | 0.00 | 0.00 | 0.00 | 0.00 | 9.91 | 0.00 | 43.69 | 337.06 | 0.00 | 0.00 | 16.00 | 445.53 |
| 1979 | 38.87 | 0.00 | 0.00 | 0.00 | 0.00 | 9.91 | 0.00 | 43.69 | 337.06 | 0.00 | 0.00 | 16.00 | 445.53 |
| 1980 | 38.87 | 0.00 | 0.00 | 0.00 | 0.00 | 9.91 | 0.00 | 69.60 | 346.97 | 0.00 | 0.00 | 16.00 | 481.35 |
| 1981 | 38.87 | 0.00 | 0.00 | 0.00 | 0.00 | 9.91 | 0.00 | 43.69 | 337.06 | 0.00 | 0.00 | 16.00 | 445.53 |

•Front end or data viewer. The last piece of the OLAP component is the front end or data viewer, and that is the component through which the user analyzes the data. Data analysis normally involves slice and dice, dimension swaping, drill down and drill thru capabilities. Among the commercial front ends we find:

–PowerPlay (Cognos)

In the graph above, an example of a multidimensional report of weather stations can be seen.

**Scorecard**

A scorecard is a common place where to monitor or give follow up to the key indicators that measure the Corporate performance. Usually each one of this indicators is compared against Budget and a tolerance is supplied for each one of them. This centralized scorecard may provide Companies an up-to-the-minute view into the health of the business. Once a deviation is found, the user can get a deeper insight by using other BI capabilities, like OLAP analysis.

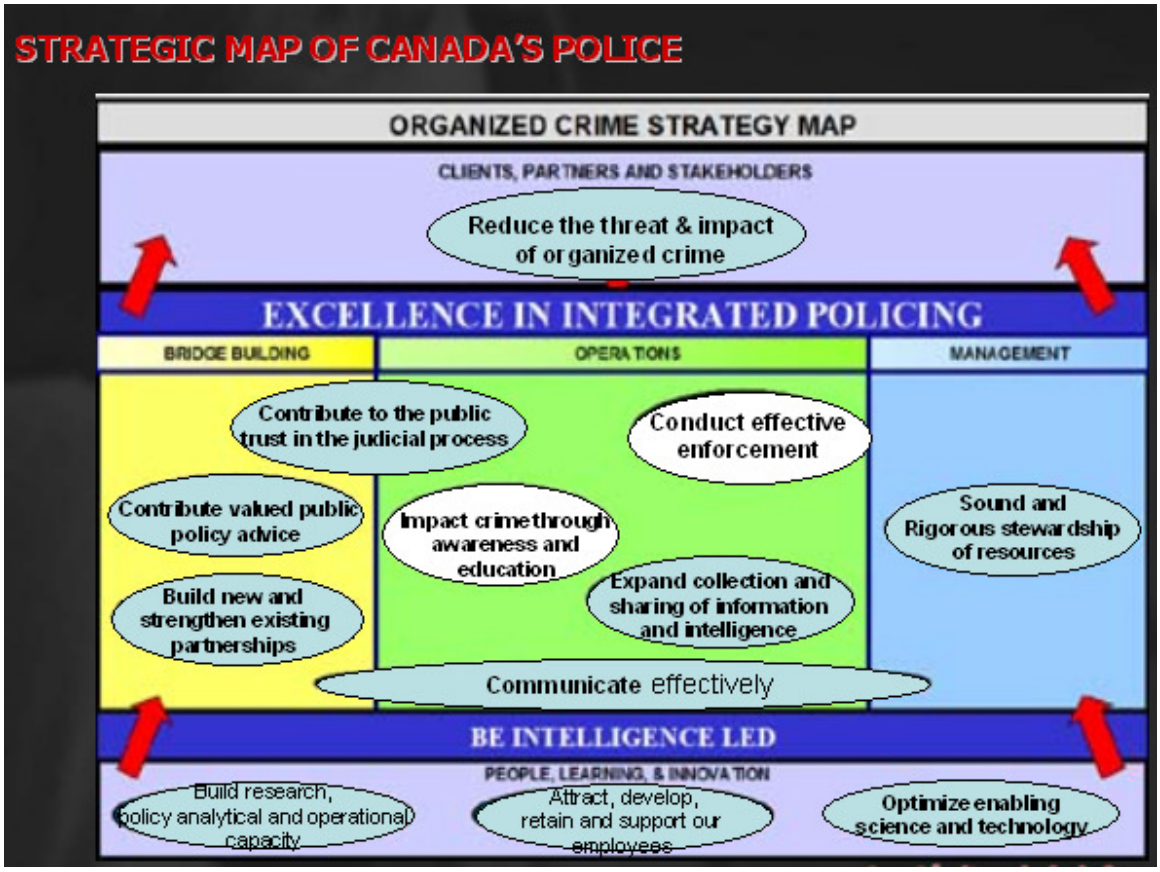
The scorecard presents operational metrics that reflect the strategical objectives of an institution. It is published so employees at all levels and in all roles understand this dashboard and are able to react against unexpected results. Each metric has a clearly identified owner, so he is in charge of triggering actions to correct its behaviour. The scorecard also brings understanding to the Metric owner about the root causes of the problem and the impact that the metric has over all the institution.

One of the benefits of this technology is that it favors early detection time, thus diminishing any negative effects. Also, if one considers the huge amount of data that is generated within the institution it is very easy to get lost as to where to start looking. In this sense a scorecard is in fact a guided visit to the relevant issues that might hinder the institution performance.

Building a Scorecard

The process begins with senior management, who define the corporate strategy, its objectives and the cause-and-effect relationships among key processes. Ownership, accountabilities, metrics and performance targets are also defined. All of these aspects are contained within a strategy map – a visual representation of how the company creates value. Distributed across the company, a strategy map shows people what the company views as important, how it wants to improve and who should be involved.

When scorecards are built and reported in isolation, there is no direct linkage to strategy. Thus, scorecards without strategy maps may lead to failure.




Strategy maps are truly like geographical maps that visually aid in understanding how does one get from A (the present capability, organization, and focus of the enterprise) to destination B (the future desired state of capabilities, organization, and focus) as laid out in the enterprise vision, mission, and strategy plan.

Each one of the strategic objectives has to be landed into operative goals. For example the objective "Reduce the threat and impact of organized crime" can be expressed in terms of number of crimes and or lost of property value. The "Sound and rigorous stewardship of resources" can be measured in terms of deviation from the original budget. The Scorecard would monitor the behaviour of each one of this indicators.

Choosing Metrics

Metrics are the building blocks of a scorecard. They indicate an area or process that the company wants to monitor, measure and improve, such as Customer Satisfaction, or Percentage of population served.

Metrics attributes:
–Target value,
–A performance threshold or tolerance,
–A clearly identified owner or responsible,
–A defined datasource that feeds it, and
–Updating frequency (latency), more frequent updating leads to smaller deviations

Each metric has an owner whose responsibility is to monitor the metric value and in case of deviations from the target take actions that would correct its behaviour. The longer the time that the metric goes out from the predesigned path the more negative impact on the whole institution.

The source or sources that provides the value or the values upon which the metric is calculated should not change in time and there has to be a fixed methodlogy according to which the metric is generated.

Metrics often map to different time frames. Generally, the more they relate to operational details, the shorter the time interval, or latency, in which they're updated. The more a metric relates to longerterm goals, the longer its latency.

For example: If the company's strategic goal is to improve the customer experience, it could monitor customer satisfaction over a quarter. It may measure the system's time down on a monthly basis, or the overall water quality could be monitored daily, which has to be expressed in the operational metric of WQI (Water quality index).

With the strategy translated into metrics, those accountable need to be able to automatically monitor them, identify issues, understand the reasons for variances and take actions to change and improve performance. This is the role of scorecarding software.

Scorecard Applications

Besides the obvious application of applying scorecard to monitor the financial aspects of an institution, like Net Income, Expenses against Income, or Human Resources aspects like headcount or Payroll, other indexes like Water Quality, Per Liter Cost, or Pumping Cost per liter could also be presented in a scorecard.

In reference to water utilities, some institutions suggest that a performance Scorecard can be created based on four broad measures: efficiency of investment; efficiency of operations and maintenance; financial sustainability, and responsiveness to customers.

Balanced Scorecard

A special case of Scorecards is the Balanced Scorecard. It is a methodology focused in the performance measurement in order to create value tor the stockholders (or society).

The Balanced Scorecard emphasizes four performance aspects:

- Finance. Institution evaluation by the investors or Credit Institutions
- Customer Relationship. Looks into how successful is the Institution in meeting customer expectations
- Internal Processes. Looks how succesful is the Institution improving the critical process of the business
- Growth and Learning. Evaluates how good is the Institution at aspects like growth, innovation and adaptability against the changing external conditions.

Within this four aspects, the Finance aspect evaluates past performance, while the other three are a look into the future. For example, the innovation capacity will derive in new patents. It is balanced because:

–Performance is evaluated through a set of balanced aspects
–Quantitative measurements equlibrate with qualitative ones (number of defects vs customer satisfaction)
–There is a balance between measurements that look into the past ( (e.g., financial ratios) and measurements that look into the future (e.g., new patents as an innovation measurement, or investment in water prospection), also between those that look into the long term with those that cover the short term.

Sometimes there is an intention to mix this four aspects into a single overall qualification, but care should be taken since it is very difficult to assign weight to each one of them. The correct way to do it would be by expressing the problem as a maximization of the Finance objective while the other aspects are expressed as constraints. In this way a shadow price could be calculated. The relation between this shadow prices can be transformed into relative weights.

**Datamining**

The objective of Datamining is to identify the relationship between a variable and a set of other variables regarded as independent. Most of the tools classified as datamining tools are based on neural networks, which is a family of algorithms that propose an equation whose parameters are slowly changed in such a way that the error between the observed target variable value and the modelled value is minimized. In order to avoid overfitting, usually this algorithms separated the data into two sets. With the first one a model is built. The second set then is run against the model in order to test it, thus constituting as an independent witness to the quality of the model built.

Datamining is considered as a Business Intelligence tool since it tries to make inferences, correlate facts, or extract "intelligence" from the same data.

Datamining algorithms are classified according to how they build a model. There are two modes: Supervised Learning and Not Supervised Learning. In the first class, the user has to specify the target variable as well as the independent variables and the possible way of processing the data. With the known outcomes, the algorithm builds a function with predictive purposes. In the second one, the user doesn't indicate the kind of relationship that exist between the different observations or events and the algorithm tries to find what are the possible subsets and what do each one of the elements of the subset has in common with the others.

Among the Supervised Learning algorithms, we have:
–Regression
–Neural Networks
–Decision Trees
–Association Rules

The Not Supervised Learning algorithms are used for Clustering/Segmentation models.

Through a search in internet, I have found the following applications of the different kinds of Datamining algorithms:

In water related issues:

- Classifiers for identification of soil types based on cone penetration tests
- Daily forecasting of river discharges
- Surge water levels prediction in the North Sea for navigation
- Text mining for document management
- Knowledge mapping in educational and engineering organizations
- Waste water treatment
- Water Reservoir Control. Datamining was used in order to obtain the best operation policy. To define an operating policy using this approach, both a single-reservoir and a multireservoir water system were modeled and optimized for a set of historical inflows. These optimization results defined the best possible performance for the systems with historical hindsight, and were used as input for the data mining process. The data mining algorithm then generated the set of control rules that gave the best historical operating policy. [5]

In other areas:

- Credit analysis
- Drugs effectiveness
- Disease precursors, Bio medical and DNA analysis
- Virus detection in computer software
- Crime Pattern Detection [1,2,3,4]

After taking a dive into all this tools, confussion might arise as when should each one of the BI tools should be used. The foolowing points could be used as a guideline:

•AD HOC Query: The user doesn't really know what he is looking for, and he is working with vast amounts of data

•OLAP: Use it to discover simple relationships between the data.

• Data mining: Prefered option to discover non obvious or hidden patterns in the data.

**Conclusions**

•BI gathers a set of technologies developed in order to take care of the good health of the institutions. It can be applied successfuly in all type of industries and in any aspect of it

•More and more datamining is going to be used in unsupected fields in innovative ways. Specially in the Water Management many documented experiences have been found that should be taken as basis for new applications

•The essential point in order to successfuly apply any Business Intelligence technology is to have clean and reliable data.

•To have a unique source of reliable data and to share it all across the corporation is of paramount importance

•A Scorecard should be used as the entry point into the BI system as a guiding tool that will take the user to the important issues. Once there, the user should have detailed reports that would allow him to understand the root causes of the deviation.

**References:**

•[1] Hsinchun Chen, Wingyan Chung, Yi Qin, Michael Chau, Jennifer Jie Xu, Gang Wang, Rong Zheng, Homa Atabakhsh, "Crime Data Mining: An Overview and Case Studies", AI Lab, University of Arizona, proceedings National Conference on Digital Government Research, 2003, available at: http://ai.bpa.arizona.edu/

•[2] Hsinchun Chen, Wingyan Chung, Yi Qin, Michael Chau, Jennifer Jie Xu, Gang Wang, Rong Zheng, Homa Atabakhsh, "Crime Data Mining: A General Framework and Some Examples", IEEE Computer Society April 2004.

•[3] C McCue, "Using Data Mining to Predict and Prevent Violent Crimes", available at: http://www.spss.com/dirvideo/richmond.htm?source=dmp age&zone=rtsidebar

•[4] Whitepaper, "Oracle's Integration Hub For Justice And Public Safety", Oracle Corp. 2004, available at: http://www.oracle.com/industries/government/Integration Hub_Justice.pdf

•[5] . J. Water Resour. Plng. and Mgmt., Volume 129, Issue 1, pp. 26-34 (January/February 2003)

Florian T. Bessler,1 Dragan A. Savic,2 and Godfrey A. Walters3

1Research Assistant, School of Engineering and Computer Science, Centre for Water Systems, Univ. of Exeter, Harrison Bldg., North Park Rd., Exeter EX4 4QF, U.K. 2Professor, School of Engineering and Computer Science, Centre for Water Systems, Univ. of Exeter, Harrison Bldg., North Park Rd., Exeter EX4 4QF, U.K. (corresponding author). 3Professor, School of Engineering and Computer Science, Centre for Water Systems, Univ. of Exeter, Harrison Bldg., North Park Rd., Exeter EX4 4QF, U.K.