APPLICATIONS IN BUSINESS INTELLIGENCE

Wolfgang ECKER-LALA*

Abstract. *In the past 10 years Business Intelligence became a more and more appropriate tool for supporting decisions. Business Intelligence Systems are based on any kind of data warehouse technology. A combination of reporting and advanced analytics creates a very efficient tool for managers. A special topic in the area of advanced analytics is datamining. Special datamining methods use e.g. Baysian methods. This will be shown in an example in which a basic rating for giving a loan will be calculated.*

Keywords: *Business Intelligence, IT systems, advanced analytics, data warehousing, reporting systems.*

1. Short Introduction in Business Intelligence

Business Intelligence (BI) is based on considerations which have been done by the Gartner Group in 1996. In fact – in the scientific/mathematical sense – there exists no real definition for BI.

In fact it has to be a tool for retrieving necessary information out of big sets of data. The implemented filtering processes has to be:

- efficient
- sufficient,

and has to provide the necessary information in an defined timeframe.

Different people have different points of view on BI. So BI can be seen as:

1. *Extension of Information Technology*

   This means that BI provides information which are extracted, aggregated etc. out of several data sources to people who has to find specific answers to specific questions based on their daily business. Using BI each query on operational data storage systems can be avoided.

* MATH-UP.COM, Landesstrasse 58, A-3441 Ranzelsdorf, Austria

2. *Filter for/against information overload*
Using BI information can be filtered. The result of such filtering are attributes which are necessary to get information in order to be able to answer specific questions. Attributes which are not necessary for finding answers will not be given to the person.

3. *Management Information System*
BI has to aggregate information out of large data sets to the minimum "point of information" which is needed to make decisions which has to be done by managers.

4. *Early Warning System*
Using appropriate information provided by BI systems can show "what is going wrong in my business?" or (which is equivalent) "was my decision which I have done some time ago correct?". If this information can be retrieved "just in time" BI can be used for managing risk.

5. *Data Warehouse*
BI stores centralized data which have been retrieved out of several operational data storage system. This is a very old definition. In fact BI is much more.

6. *Storage of Information and Knowledge*
BI uses data out of several operational systems, combines it with data out of (maybe external) databases and provides us data which can be results out of "artificial intelligence". So we can get new knowledge and we can retrieve information of events which have been done or seen or investigated in the past.

7. *Process*
BI is a set of tools or a system which allows us to do some investtigation of symptoms, make diagnosis, decide the "correct" therapy, gives us the possibility to make a prognosis (if necessary) and is a tool for controlling if our chosen therapy is opportune.

In fact BI is:

- *Integrated*
  which means that it is integrated into all systems of a company. It is not a stand-alone system because it needs data out of all available operational systems and it has to provide data and results to reporting and decision systems of a company or institution.

- *Specific to an enterprise/company*
  which means that BI has to fulfil all requirements of information needs which are requested by decision makers of a company/enterprise and therefore it has to be specific to each company.
- *IT-based*
  because of large sets of information and the requirement of efficiency and sufficiency it has to be based on IT-systems.

BI is an overall attempt to get a base for decision at operational level. and it is more than:

- storing data in a database
- creating reports
- finding errors or inconsistency within data.

Business Intelligence Systems have to support the management of a company in taking decisions and provide with all necessary reports (Fig. 1).

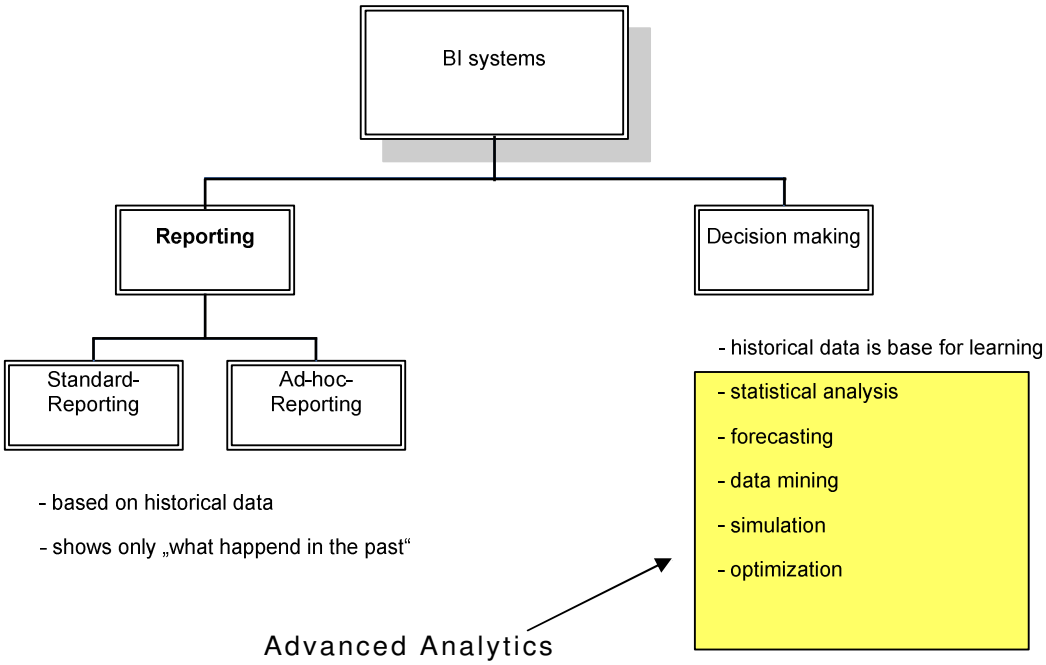

Figure 1. BI Systems function diagram.

So this paper will show how business intelligence can be enhanced by statistical methods in order to transform information into knowledge.

2. Reporting

A report provides an overview of a well defined topic and is based on historical data. Visualisation techniques like diagrams and graphs improve the way of understanding of the information transported by the report to the receiver.

In many cases a reporting system is required by a national regulator, e.g. in banks, insurance companies, companies noted on the stock exchange.

There are several processes within reporting as shown in the graphic below (Fig. 2).

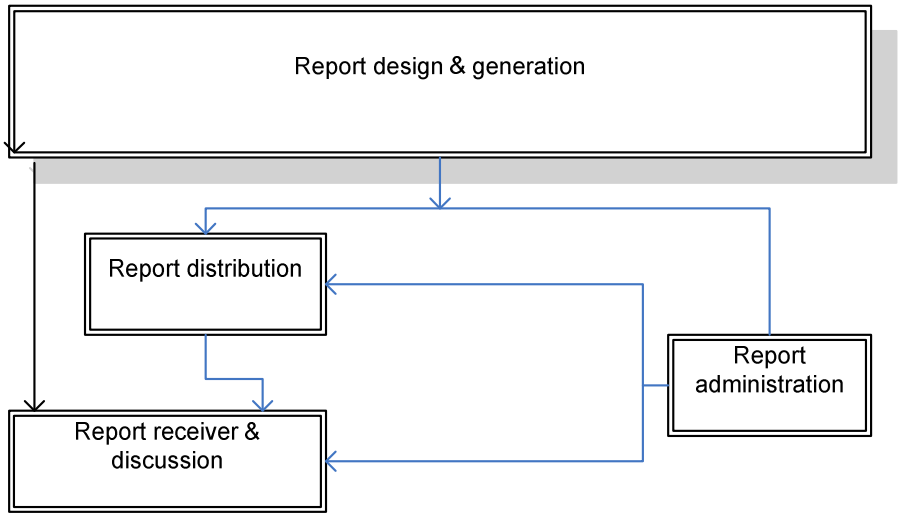

Figure 2. The processes within reporting system.

Report design and generation

In the process of "report design and generation" layout and content of the report are defined and fixed. The content of the report (and of course the layout) has to fit to the needs of the receiver.

Report distribution

In this process the receiver of the report and the frequency of the report delivery to the receiver are defined. At least even the mode of delivery (e.g. e-mail, EXCEL sheet, …) is defined here.

Report administration

This process takes care of the quality assurance and the availability of the report. Also the access rights to the report are defined and administered

in this process. And last but not least it has to take care about accessibility and availability of historical reports.

Report receiver and discussion

This process handles the receiving and the interpretation of the delivered information. And of course it should start discussions among the report's receivers.

Standard-/Ad-hoc reporting

These two subsystems are categorized into

Active

Reports of this subsystem are generated periodically by the system. This means that the event to generate a report is triggered by the system itself. This subsystem is called ***Standard-Reporting***.

Passive

Generation of reports of this subsystem is triggered by a person who needs specific information at a non-predefined time. This subsystem is called:

Ad-hoc-Reporting

Following graphic shows that a reporting system consists of two subsystems (Fig. 3).

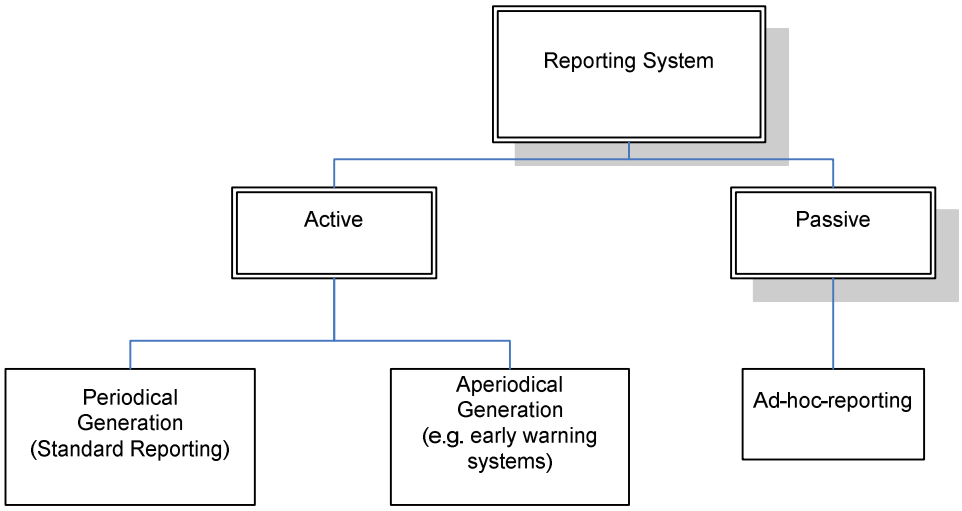

Figure 3. Structure of Reporting Systems.

3. Advanced Analytics

During the last 10 years there has been an explosion of systems of information technology. We are collecting a huge amount of data every

day. These data are used in a variety of fields such as: medicine, biology, finance, marketing, controlling etc. and new areas such as data mining, machine learning and bio informatics has been developed.

In the past those data have been collected for reporting purposes. Up to now reporting means "looking back into the past".

We think that these times are gone and we have to spend money – maybe more money as we like to spend now – to analyse data of the past to be able to predict future events in order to be able to get better decisions.

Therefore all classical methods of statistics and forecasting have to be combined with traditional reporting techniques in Business Intelligence Systems of the future. Let us call this new generation of Business Intelligence "Advanced Analytics".

Maybe the current "economic crises" could have been avoided if more advanced analytical methods had been used in monitoring risks.

3.1. We have a lot of data – and now?

Most of the companies store a lot of data out of their daily business. These data are used for creating reports and finding information when they are running queries on their databases. But all those reports and query results provide data which gives us information of the past.

On the other hand companies spend a lot of money to be able to store all the data. These money is spent for the:

- physical storage systems
- operational handling of these storage systems
- backup and recovery handling of these storage systems
- availability of the data.

So if we consider all these efforts in order to be able to answer the questions:

- Did Mrs. Miller buy a new car 5 years ago?
- Did she like a red or a blue or a green car?
- Did we sell 500.000 cars to women during the last 5 years?

These questions raise up if all these money which has been spent is really a **R**eturn **O**f **I**nvestment for this information.

Is it really necessary if a company or institution retrieves information out of expensive information systems in order to see what was done in the past and how much profit could be got in the past or is it more important to be able to predict what will be in the future? Is it satisfying if databases are not more than **graveyards of data**?

3.2. Requirements

Regarding the considerations we did the reality is as follows. In fact a company needs:

- *historical data*
  in order to be able to store all the knowledge which could be retrieved in the past;
- *"clean" data*
  in order to be able making good decision for future activities and to become better and in order to be able to report consistent results of the past;
- *a workaround for missing data*
  in order to be able to report if there is something missing in the stored information and to be able to substitute the missing information;
- *aggregated data*
  in order to be able to do calculations and provide results within a timeframe which can be accepted by a decision maker;
- *base of information for making decisions*
  in order to have reliable data where the decisions are based on:
- *algorithms which gives support in decision making,*
  in order to be able to combine scientific knowledge and available data.

All these can be done using a methodology which is called Data Warehousing. This methodology will be explained in the next section.

3.3. Data Warehousing

A Data Warehouse is a dispositive data storage system which is separated from all operative databases and data files.

The architecture of a data warehouse depends on the used technology and has to be designed to retrieve information in a very short time using a huge size of data.

The data warehouse database(s) collect(s) information out of several separated data sources and so there must be a lot of interfaces provided. The process to get data into a data warehouse is called ETL process (**E**xtract – **T**ransform – **L**oad). After this process is finished data has to be cleaned in order to get a "unique truth". All information after the data cleaning has to be without contradiction.

The attributes of a data warehouse are:

- *Subject oriented,*
  which means that such kind of system are always based on the needs of management of a company.
- *Integrated,*
  which means that a data warehouse has to be able to work together with all other data storage systems in a company.
- *Referenced to time periods*,
  which means that all the information which were imported by ETL are dependend on a time based snapshot (e.g. every day, every week, …).
- *Non-volatile,*
  which means that the values of a data record will not change over time.

  In general we distinguish two architectures of data warehouse:
- *Centralized,*
  is physically based on one server system.
- *Decentralized,*
  which means that more than one physical data warehouse exist. Here it can be that the information is spread over the data warehouse systems and all these systems together shows the complete information e.g. of a company.

4. Advanced Analytics in a banking application

After all this theory and lot of ideas it is time for showing a real application of "advanced analytics".

In the daily business a bank has to consider the risk of a customer who asks for a credit or loan. So following problem has to be solved.

The bank has to know if a customer who asks for a credit has to be categorized as a potential defaulted borrower or not.

This problem should be solved by developing a data mining algorithm. Therefore based of the existing historical data we will need a training set and a evaluation set.

Let us assume that the data set which will be the base of the decision has following attributes:

- Home owner
- Marital status
- Approx. annual income.

So if we consider each record as a stochastic variable X and the class (default or no default) as stochastic variable Y we have to calculate if:

$$P(Y = \text{yes} \mid X) < P(Y = \text{no} \mid X)$$

or:

$$P(Y = \text{yes} \mid X) > P(Y = \text{no} \mid X)$$

and based on this the bank has a base for the decision of giving the loan to the customer or denying it.

In order to be able to develop a good data mining algorithm we have to take a sample of the existing records as a training set and we will do following categorization:

- *Class = YES*,
  if loan owner has defaulted of her/his payments;
- *Class = NO*,
  if loan owner has repaid the complete loan,

e.g. the training set looks like in Table 1.

Table 1

ID	Home Owner	Marital Status	Annual Income	Class
1	Yes	Single	125K	NO
2	No	Married	100K	NO
3	No	Single	70K	NO
4	Yes	Married	120K	NO
5	No	Divorced	95K	YES
6	No	Married	60K	NO
7	Yes	Divorced	220K	NO
8	No	Single	85K	YES
9	No	Married	75K	NO
10	No	Single	90K	YES

In this example we have two categorical attributes:

- Home owner (= X_1),
- Marital status (= X_2)

and one attribute which is continous,

- Approx. annual income (= X_3).

If we choose a Bayesian approach we have following relationships between prior, conditional and joint probability functions:

$$P(X,Y) = P(X \mid Y) \cdot P(Y)$$

and:

$$P(X,Y) = P(Y \mid X) \cdot P(X).$$

From this we get:

$$P(Y \mid X) = \frac{P(X \mid Y) \cdot P(Y)}{P(X)}.$$

If we select n independent observations from the historical $\{X_1, X_2, ..., X_n\}$ data we can calculate:

$$P(X \mid Y = y) = \prod_{i=1}^{n} P(X_i \mid Y = y).$$

And it is easy to understand that we derive:

$$P(Y = y \mid X) = \frac{P(Y = y) \cdot \prod_{i=1}^{n} P(X_i \mid Y = y)}{P(X)}.$$

So we have to consider this for the categorical attributes and get for:

- Home owner (= X_1);
- Marital status (= X_2):

$$P(X_i = x_i \mid Y = y) = \frac{n_i}{m_i}$$

where:

n_i is number of records with value (x_i, y);

m_i total number of records having y.

And based on the previous formulas we get for the categorical attribute "Home owner":

$$P(X_1 = Yes \mid Class = NO) = \frac{3}{7}$$

$$P(X_1 = No \mid Class = NO) = \frac{4}{7}$$

$$P(X_1 = Yes \mid Class = YES) = 0$$

$$P(X_1 = No \mid Class = YES) = 1.$$

Based on the previous formulas we get for the categorical attribute "Marital status":

$$P(X_2 = Single \mid Class = NO) = \frac{2}{7}$$

$$P(X_2 = Divorced \mid Class = NO) = \frac{1}{7}$$

$$P(X_2 = Married \mid Class = NO) = \frac{4}{7}$$

$$P(X_2 = Single \mid Class = YES) = \frac{2}{3}$$

$$P(X_2 = Divorced \mid Class = YES) = \frac{1}{3}$$

$$P(X_2 = Married \mid Class = YES) = 0.$$

Last but not least we choose for the continuous attribute:

- Approx. annual income (=X_3)

a Gaussian distribution.

If we take following consideration into account:

$$P(x_i < X_i < x_i + \varepsilon \mid Y = y_i) = \int_{x_i}^{x_i+\varepsilon} f(X_i \mu_{ij}, \sigma_{ij}) dX_i \approx f(X_i \mu_{ij}, \sigma_{ij}) \cdot \varepsilon$$

we get:

$$P(X_i = x_i \mid Y = y_j) = \frac{1}{\sqrt{2 \cdot \pi} \cdot \sigma_{ij}} \cdot e^{-\frac{(x_i - \mu_{ij})^2}{2 \cdot \sigma_{ij}^2}}.$$

The estimation of the parameters for this distribution is:

- in Case of Class = *NO*:

$$\mu_{ij} = 110000, \sigma_{ij}^2 = 2975$$

- in Case of Class = *YES*:

$$\mu_{ij} = 90000, \sigma_{ij}^2 = 25.$$

If e.g. a Customer has the attributes:

- Home Owner = *No*
- Marital status = *Married*
- Approx. annual Income = 120.000;

we get following:

$$P(X \mid Class = NO) =$$

$$= P(X_1 = No \mid Class = NO) \cdot P(X_2 = Married \mid Class = NO) \cdot$$

$$\cdot P(120000 \mid Class = NO) = \frac{4}{7} \cdot \frac{4}{7} \cdot 0.0072 = 0.0024$$

and:

$$P(X \mid Class = YES) =$$

$$= P(X_1 = No \mid Class = YES) \cdot P(X_2 = Married \mid Class = YES) \cdot$$

$$\cdot P(120000 \mid Class = YES) = 1 \cdot 0 \cdot 1.2 \cdot 10^{-9} = 0$$

and:

$$P(Class = YES) = \frac{3}{10}$$

$$P(Class = NO) = \frac{7}{10}.$$

From all above we can calculate:

$$P(Class = YES \mid X) = P(X \mid Class = YES) \cdot P(Class = YES) = 0 \cdot \frac{3}{10} = 0$$

and:

$$P(Class = NO \mid X) =$$

$$= P(X \mid Class = NO) \cdot P(Class = NO) = 0.0024 \cdot \frac{7}{10} > 0.$$

And so in this example the bank decides that the customer will get the loan because the probability that he/she will do the complete repayment is greater than he/she will be defaulted.

Of course this example is a very simple one but it shows how statistical methods have to be combined with information which can be found in a data warehouse and can be derived from a huge amount of data records.

This paper should animate the reader to think about how a BI system has to be designed to get an efficient data base for a lot of decisions which have to be done in the daily business. It would be a very good exercise to design the part of a business intelligence system which will be necessary to solve the bank's decision problem which was described.

5. Conclusions

The usage of business intelligence tools offers a wide spectrum for controlling, forecasting and decision support to the management of a company. Combined with mathematical and statistical methods IT systems will become much more efficient.

Data are the "gold" of a company. We are living in a world of a lot of information. Humans changed from *homo sapiens* to *homo informaticus*. So we and our IT systems are asked to use data in a much more efficient way as it was done in the past 20 years.

REFERENCES

[1] Cord Spreckelsen, Klaus Spitzer, *Wissensbasen und Expertensysteme in der Medizin*, Vieweg + Teubner, ISBN 978-3-8351-0251-4 (2008).

[2] Peter Zische, *Business Intelligence für kleine Unternehmen*, W3L-Verlag, ISBN 3-937137-51-3 (2004).

[3] Hans-Georg Kemper, Walid Mehanna, Carsten Unger, *Business Intelligence – Grundlagen und praktische Anwendung*, 1. Auflage, Vieweg, ISBN 3-528-05802-1 (2004).

[4] Pang-Ning Tan, Michael Steinbach, Vipin *Kumar*, *Introduction to Data Mining*, Addison Wesley, ISBN 0-321-32136-7 (2006).

[5] David Hand, Heikki Mannila, Padhraic Smyth, Principles of Data Mining, The MIT Press, ISBN 0-262-08290-X (2001).